Recombination and the Structure of Globular Proteins

Diplomarbeit

zur Erlangung des akademischen Grades

Magister rerum naturalium

an der
Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

Vorgelegt von

Jörg Hackermüller

im Februar 2001

meinen Eltern

nicht nur dafür, dass sie diese Arbeit ermöglicht haben

Dank an alle,

die zum Gelingen dieser Arbeit beigetragen haben:

Peter Stadler, Ivo Hofacker, Peter Schuster.
Ingrid Abfalter, Aderonke Babajide, Jan Cupal, Martin Fekete, Christoph Flamm, Dagmar Friede, Kurt Grünberger, Christian Haslinger, Philipp Kobel, Michael Kospach, Ulli Mückstein, Stefan Müller, Bärbel Stadler, Roman Stocsits, Andreas Svrcek-Seiler, Caroline Thurner, Günther Weberndorfer, Andreas Wernitznig, Stefanie Widder, Christina Witwer, Michael Wolfinger, Daniela Dorigoni, Judith Jakubetz.
Anne-Kartrin Neyer, Lucia Hackermüller, Peter Kolb, Andreas Hochwagen, Angelika Küng, Florian Triebl, Josef Diewok, Andreas Lämmerhirt, Arash Pourkarami.
Christina Hackermüller, Alois Hackermüller, Gertraude Rotter, Franz Rotter, Helmut Pschorn.

Zusammenfassung

Rekombination, ein Mechanismus um genetische Information zu mischen, scheint ein vollständig aufgeklärtes genetisches Phänomen zu sein. Üblicherweise sieht man in der Rekombination einen evolvierten Mechanismus, dessen Entwicklung zu einer Verbesserung und Beschleunigung des Evolutionsprozesses geführt hat. In dieser traditionellen Vorstellung wird die genetische Stabilität von Individuen durch die hohe Genauigkeit der Replikation gewährleistet, während Rekombination für ausreichend Variabilität innerhalb der Population sorgt, indem sie allfällige Mutanten zwischen den Individuen verteilt und sie dadurch in einen neuen genetischen Kontext bringt. Diese Vorstellung beruht auf der Sicht einer überwiegend vorteilhaften Rekombination, man erwartet Rekombinationsprodukte, die in der Regel überlebensfähiger sind als ihre Eltern.

Im Widerspruch zu dieser Vorstellung, konnten wir in dieser Arbeit zeigen, dass Rekombiation zwischen nicht homologen Genen, die für den gleichen Genotyp kodieren, die Struktur in der Mehrheit der Fälle zerstört. Wir prüfen daher die Annahme, dass es sich bei Rekombination nicht um eine evolvierte Fähigkeit von Sequenzen handelt, sondern um einen unvermeidbaren Effekt der Nukleinsäurechemie. Natürlich müssen sich nicht alle Rekombinationen negativ auswirken. Rekombinationen an Positionen, die funktionell oder strukturell selbständige Einheiten trennen, sollten weniger negative oder sogar positive Auswirkungen haben.

Eine mögliche Strategie von Genen oder ganzen Genomen um auf die zerstörerische Rekombination zu reagieren ist, die Abstände zwischen Loci abhängig von der Stärke ihrer epistatischen Wechselwirkung zu variieren. Das Inserieren von Introns in Positionen wo eine Rekombination harmlos ist, könnte eine Möglichkeit sein, um die Wahrscheinlichkeit einer Trennung durch Rekombination zwischen interagierenden Loci zu senken und Rekombination auf Bereiche zu konzentrieren, wo wenig Schaden zu erwarten ist. Eine Untersuchung von Introns in Regionen die zwischen Eukaryonten und Prokaryonten konserviert sind, ergab, dass Introns an Modulgrenzen gehäuft auftreten. In dieser Arbeit untersuchen wir die Verteilung von Introns mittels Computersimulation eine Flussreaktors. Eine signifikante Korrelation zwischen der Struktur und der Intronverteilung konnten wir nicht feststellen. Rekombination verursacht unter den Bedingungen dieser Computersimulation offensichtlich keine Modularisierung von Genen globulärer Proteine in ihre funk-

tionellen oder strukturellen Bausteine.

Offensichtlich kann eine Computersimulation die experimentellen Ergebnisse nicht reproduzieren. Die Gründe dafür sind entweder in einem ungeignetem Aufbau der Computersimulation, die eine vorhandene Tendenz zur Modularisierung nicht detektieren könnte oder in den verwendeten Fitnessfunktionen zu suchen. Möglicherweise unterstützen die von *Knowledge Based Potentials* abgeleiteten Fitnessfunktionen eine Fragmentierung von Genen nicht. Dies kann durch eine algebraische Analyse der durch die Fitnessfunktionen bestimmten Landschaften abgeklärt werden, wofür derzeit aber keine etablierte Technik zur Verfügung steht.

Abstract

Recombination – a mechanism of mixing genetic information – seems to be well understood. Recombination is usually seen as a process, which evolved to streamline the process of evolution. In this picture, high fidelity replication accounts for the genetic stability of an individual, whereas recombination creates some variability by mixing mutated products of replication and bringing them into a new genetic context.

In this work, however, we show that recombination among non-homologous genes, which code for the same RNA secondary structure, or for amino acid sequences folding into the same globular protein structure, disrupts the structure in the majority of cases. We therefore investigate the assumption that recombination is not an evolved beneficial mechanism, but an unavoidable side effect of nucleic acid chemistry. Of course not all recombination events must have a negative effect. Recombination at positions which separate functionally self-contained units should be less negative or even beneficial for the fitness of the offspring created.

One possible response of genes or the entire genome to the negative effects of recombination is to modify the distances between loci depending on the strength of their epistatic interactions. In other words the insertion of introns into positions where recombination does not do any harm could have been a means to keep the probability of recombinational separation of two interacting loci low while directing recombination to harmless positions. An investigation of the intron positions in ancient conserved regions of genes revealed that a part of the introns, is indeed located at module boundaries. We studied the placement of introns under recombination and point mutation in an evolutionary flow reactor simulation. We do not find a significant correlation between the placement of introns and the protein structure. Recombination therefore does not cause a modularisation of the genes of globular proteins into their functional or structural building blocks under the conditions of computer simulation presented here.

Consequently we have to ask why a computer simulation is not able to reproduce the experimental results. Either our experimental setup is not suited to detect such module boundaries, or the fitness functions we derived from knowledge-based potentials do not support the modularisation of genes.

Contents

1 Introduction

Recombination – a mechanism of mixing genetic information – seems to be well understood. Looking up "recombination" in standard text books of molecular biology [1] or genetics leaves no doubt that recombination evolved over aeons to streamline the process of evolution. High fidelity replication with rare point mutations accounts for the genetic stability of an individual, whereas recombination creates a certain variability by mixing the possibly mutated products of replication bringing them into a new genetic context.

Publications on the origin of recombination [8, 31, 73, 22, 41], however, indicate that at least this question, which is usually not mentioned in textbooks, is still a matter of argument. Albeit there are some differences about how recombination evolved (see [37] for a classification of hypothesis for the evolution of genetic mixis), there is a strong but startling consensus that recombination is an evolved feature. More precisely, recombination is viewed as a feature that has been acquired by nucleic acid sequences during evolution. Of course one could ask why a process that almost all present-time natural sequences are capable of, should have been acquired in the course of evolution rather than considering it as an unavoidable process in nucleic acid chemistry. The main argument for an evolutionary origin appears to be the perception that recombination is always beneficial. In other words recombination provided such a large advantage, that only "recombinable" sequences survived.

C. Biebricher and R. Luce utilized a system originally developed by I. Haruna, R. Levisohn and S. Spiegelman [29, 62] to study molecular evolution *in vitro* [5, 4]. They used a simple RNA polymerase, from the phage $Q\beta$, which is able to amplify all RNA sequences that form a particular secondary structure element. The replication of a highly optimized RNA species yielded not only complementary copies, but also longer RNA molecules which were not able to replicate. Biebricher et al. explained the formation of the longer replication product by a recombination mechanisms during replication. Munishkin [51] found several short by-products in an analogous experiment which he related to recombination events. Biebricher [5] explained the formation of the recombinants by a mechanism, where the polymerase was simply falling off the template chain, eventually continuing on a different template or on the same chain at a different position. Obviously, this form of recombination requires no complicated apparatus and no special relationship between the sequences combined by the recombination event. In contradiction to the assumptions about recombination above, this basic form of recombination is only rarely beneficial, as it mainly leads to a combination of unrelated sequence fragments. This raises the question, whether recombination is really an acquired property or rather an accident, happening to all sequences since the invention of replication.

Evolution then would have tamed the ubiquitous recombination and restricted it to its beneficial present day forms. Genes and genomes in this picture have always been under the effect of recombination, which originally was detrimental in almost all cases and have adapted to it. Today's textbook forms of recombination are the outcome of this adaption: recombination has been restricted the control of an elaborate molecular apparatus, to act only on related sequences, therby largely avoiding the negative effects.

Some support for this hypothesis comes from the theory of genetic algorithms. Genetic algorithms, which utilize recombination as a move operator, were invented to solve optimization problems on rugged landscapes. A comparison of the performance with other optimization methods like simple hill climbing revealed, that they are often outperformed, even on landscapes constructed as paradigms for the demonstration of genetic algorithms [49, 34]. The effort made currently to find adapted forms of genetic algorithms and appropriate landscapes [33], to demonstrate their supremacy over different evolutionary computation approaches in finding the global optimum of a landscape, indicates that recombination does not at all guarantee a better optimization performance in general. In fact the "No Free-Lunch" theorems [71] show that there is no algorithm taht is superior on all landscapes.

A change in the view about the origin of recombination does not leave other important biological problems completely untouched. Recombination plays a crucial role in the theory of how intergenic sequences evolved. At the current stage of sequencing the human genome, it is estimated that only 5-10% of the human genome carry essential information. Genes are not only separated by large stretches of nonsense DNA but also interrupted by non-coding sequences called introns [14, 11], which have no corresponding in the mRNA and the protein. The origin of introns is still an issue of scientific discussion [54, 16]. Two competing theories assume that introns either have been always there [24, 26, 57] or developed rather late in evolution by random insertion [14, 58, 44].The hypothesis of "tamed recombination" does not imply either one of the two hypotheses, but rather suggests a new "introns intermediate" hypothesis. Introns could have been a means of adaption in response to recombination.

A recombination event within two intron regions should not do any harm, because the region around the recombination position is not translated into the protein anyway. This should direct intron development and select for introns in positions where a recombination event is harmless. Recombination is beneficial if it recombines modules which have strong epistasis within and weak coupling between them. Therefore we should observe introns in locations that separate genomic "modules". This model does not require a special recombination hot-spot in the intron sequence, but relies on the fact that the probability to separate two loci by recombination is proportional to the distance between the loci.

Prior to studies about the effect of recombination on intron development, we will test the assumptions made so far about the role of recombination in evolution. So far, there is no model that properly explains epistatic interactions in entire genomes. We therefore restrict ourselves to the investigation of the impact of recombination on "genomes" that contain only a single gene. We will investigate the effect of a recombination event by considering populations of haploid species with RNA or DNA genomes which are able to reproduce only via recombination. A fitness function which describes the ability of an individual to fulfill a certain biological function, gives us a tool to compare populations in terms of survivability. The impact of recombination at a certain position can be estimated by calculating the mean fitness difference over all recombinants, obtained from the parental population. If the assumption is true that recombination is disadvantageous in the majority of cases, one should expect that the average fitness of the offspring decreases with the distance of the recombination point from the ends of the sequences, because the exchange of a short subsequence in the terminal region will not effect the function of the biopolymer as much as in the core region. A recombination event at a position which separates self-contained functional elements (which we consider equivalent to structural) should have a less negative or even positive impact on the survivability of the offspring compared to recombination at an arbitrary position. Such cross-overs should yield a better mean fitness difference, thereby forming an exception from the general trend.

The mean fitness difference between recombinant and parents over all recombinants, is a reasonable measure for the impact of recombination, that should allow to identify positions that are less than average susceptible to cross-over events. In an evolutionary system, however, not the mean fitness produced by a crossover event at a certain position is important, but the number of offsprings with a better fitness than their parents. Even a recombination producing only a single child with higher survivability, which could give rise to a new clone, should be called a beneficial event. Therefore we shall additionally calculate the fitness distribution in the offspring and the fitness of the best offspring produced.

The effect of recombination in an evolving system can be studied in more detail by a genetic algorithm, which produces offspring the like as above, generating all $n \cdot (n-1)$ recombinants per position. The n parents of the next generation are chosen by calculating the best recombinants per crossover position and selecting the n fittest out of this group. Again we are interested in the best and the mean fitness differences between offspring and parents. In its traditional view, recombination is thought to create variability in the population. Based on a formal investigation of recombination [64, 68, 65] we expect – in contradiction to the traditional view – recombination to act as a homogenizing operator. The changes in the variability of the offspring generation compared to the parental population can be detected with the genetic algorithm in the following reproduction cycle. If recombination truly homogenizes the population, we are able to study the effects of recombination in a more and more homologous genetic environment.

How does recombination affect the placement of introns in the genome? The probability to separate two loci on the genome by crossover depends on the physical distance between the loci. Loci adjacent to each other are rarely separated by recombination, the recombination frequency between two loci increases with their distance. The upper limit of the recombination probability between two distant loci is given by the case of two loci situated on two different chromosomes. The probability to recombine the chromosomes, i.e. to select a maternal and a paternal instead of two maternal or two paternal ones is 50%. Two very distant loci located on the same chromosome behave, in terms of recombination frequency, like loci on different chromosomes. The placement of introns between two loci obviously alters the probability to separate the loci via crossover. If one assumes that the recombination frequency is independent of the total genome length, the placement of an intron between a pair of loci will increase the recombination probability between the loci, but reduce the probability of separation between any other pair of (not too distant) loci. Consequently one would expect to find introns preferentially between genes and between functionally self-contained regions of genes. A recombination event between epistatically interacting loci, e.g between loci which participate in one alpha helix in the protein is usually disadvantageous. Therefore one would expect two such loci to be located as close as possible to each other and not to be interrupted by introns.

We are going to test the hypothesis about the modularization of protein genes under the pressure of recombination in a tank reactor simulation. The reactor is filled with an initial population of species with haploid DNA genomes, which encode a single protein. The phenotype of an individual is the possibility of the encoded amino acid sequence to fold into a given protein structure. In the initial population the phenotypes are roughly equal. For each individual the probability to reproduce depends on its fitness. The probability to die is equal for all species and depends only on the number of species in the reactor, to keep the number of individuals in the reactor approximately constant. During replication the individuals are susceptible to a series of genetic operations: point mutation, recombination and the insertion and deletion of introns. Introns are symbolized by special intron characters different from the coding letters to simplify the simulation of splicing procedure prior to translation. The evolution of introns is studied by calculating the intron distribution per sequence position for various reactor time steps. To judge whether introns form preferentially between self-contained structural units, we associate each codon in the DNA with the secondary structure in the corresponding protein and calculate the fraction of introns, which have formed inside and between secondary structure units. Obviously we would expect to find the highest fraction of introns between secondary structure elements, a certain amount in regions of hardly defined secondary structure like coil or turn regions and only a small fraction of introns within secondary structure elements, like alpha helices or beta sheets.

Additionally we can use the tank reactor to gain further knowledge about the performance of recombination in an optimization problem. The fitness averaged over all individuals in the reactor is a good means to compare reactor runs with point mutation and recombination and with point mutation only.

2 Theory

2.1 Biological Aspects of Recombination

Recombination is a genetic process which is mixing the information between two nucleic acid chains. This mixture occurs by cutting both strands into two pieces and religating them crosswise. Figure 1 illustrates what is meant by a recombination involving two DNA strands. Usually a second term is used for this process called crossover. A crossover often describes rather the outcome of recombination and not the recombination process itself. However for this work "recombination" and "cross-over" are used interchangeably. In the field of genetics the term "recombination" is usually used for a trivial process of genetic mixture as well. In diploid sexually reproducing organisms the maternal and paternal alleles residing on different chromosomes are mixed by the fair transsion of chromosomes into gametes, a process called meiosis. This process of genetic mixture will not be considered any further in the course of this work. Unless stated otherwise, the term "recombination" always implies the breakage and reunion of DNA strands. Figure 1 shows a recombination where only one double strand break and re-ligation occurs. This is called a one-point crossover. In principle several cuts and religations happen during a crossover, however here we restrict ourselves to one-point crossover or recombination. Until further notice the terms recombination and cross-over actually mean one-point cross-over and one-point recombination.

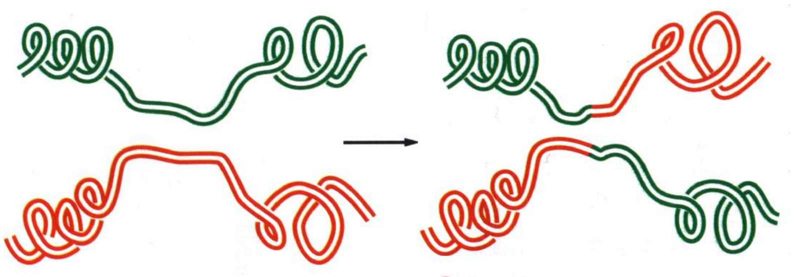


Figure 1: A pictorial description of what is meant by a recombination between two nucleotide sequences

Recombination is found in several forms in nature. The homologous or general recombination requires sequence identity or high homology between both sequences around the cross-over position. Site specific recombination has no homology constraints but requires a consensus sequence in one of the recombination partners. Transpositional recombination is used to insert mobile genetic elements and needs a minimal homology between both sequences. A fourth form of recombination is usually not mentioned in textbooks. It is the cross-over outlined already in the introduction, which shall be called simple recombination.

2.1.1 General Recombination

General Recombination is maybe the most elaborate form of recombination. Furthermore, it is the most important type of recombination for present-day organisms. All sexually reproducing diploid organisms are dependent on general recombination. The reduction from a diploid cell to haploid gametes implies not only a genetic mixing by creating a mixed set of maternal and paternal chromosomes, but also an exchange of genes or parts of genes between widely homologue chromomes, which is achieved by general recombination. This form of recombination plays also a major role in the exchange of genetic information among bacteria by a mechanism called conjugation. Obviously general recombination must have been invented before any form of sexual reproduction took place. Most studies on general recombination were performed with *E. coli.*

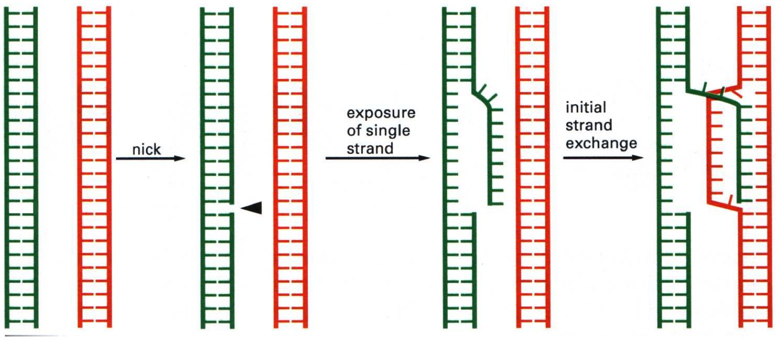


Figure 2: The mechanism of homologous (general) recombination

General recombination is dependent to a large extent on homology between the two participating sequences. Considering the mechanism of general recombination [42] this can be easily understood. Figure 2 illustrates this mechanism. This type of recombination requires proximity between both DNA helices. Furthermore the free single strand must be able to form stabile base-pairs with the second DNA molecule to displace the originally pairing strand. In E. coli a protein and a protein complex are necessary for this process: RecBCD enters the DNA, forms a nick and displaces a single stranded “whisker”. RecA is necessary to stabilize the single stranded DNA and to allow the displacement of the base-pairing strand in the second DNA molecule [56]. Though homology seems to be required for general recombination, in vitro experiments on the role of recA showed that recombination can take place when the homology is not perfect. However it is assumed that the cell’s DNA proof-reading system prevents promiscuous recombination. Closely related genes of *E. coli* and *Salmonella typhimurium* will not recombine though they have 80% homology. When the proof-reading system is knocked out, both sequences recombine easily [55].

The complex process and the requirement of specialized proteins suggests that general recombination was invented rather late in evolution. At least later than the time when the first primitive proteins could have been patchworked together by intronic recombination as it is claimed by the introns late hypotheses discussed in chapter 2.3. Moreover a consensus sequence among introns does not exist and the chance that introns show sufficient homology to allow a general recombination event is low. Consequently this type of recombination should not have played a role in the evolution of introns.

2.1.2 Site Specific and Transpositional Recombination

Site specific recombination is performed by a recombination enzyme that recognizes specific nucleotide sequences on at least one of the recombining DNA molecules. Homology is not necessary for this form of recombination, if heteroduplexes between both DNA molecules are formed at all, they are usually only a few nucleotides long. Site specific recombination is a means for bacteriophages and retroviruses to enter their host’s genomes. Mobile DNA sequences use this process to move around within and between chromosomes.

Sequences utilizing site specific recombination use the same recombination enzyme, encoded by integrated sequence, to leave the host DNA molecule, which happens often under stress e.g. in bacteriophage lambda [40]. These excision process restores exactly the two original molecules, therefore this form of recombination is called conservative site specific recombination.

Transposable elements use a somewhat different form of recombination. They do not need a specific signal on the host DNA, but simply break two phosphodiester bonds, resulting in a cut double strand with overlapping sticky ends. The transposable element invades the host DNA by forming short heteroduplexes, which are usually shorter than the overlap, resulting in a gap of unpaired host DNA [50]. These gaps are filled by DNA repair, thereby generating flanking duplications of the host DNA. Because of this duplications, transposable insertions can be easily identified.

Integration of transposable elements by recombination is thought to be the cause for some more recent intergenic sequences. The flanking duplications make transposable introns distinguishable from ancient introns. Unlike introns accounting for exon-shuffling, introns generated by site specific or transposable recombination are not restricted to a specific intron phase.

2.1.3 RNA recombination

RNA recombination has been discovered already in the early days of molecular biology, however only in the recent history various forms of RNA recombination could be distinguished and mechanisms could be proposed. RNA recombination was found the first time as an exchange of genetic information between closely related RNA viruses infecting the same cell. Unlike DNA recombination in its forms described above no specific genes could be found and it was was not possible to dissect the mechanism by genetic means. A biochemical analysis was not possible those days, because a cell free system for RNA recombination had not been developed yet. So it could not be elucidated whether this form of crossover was promoted by host or viral proteins. Moreover it was unclear whether it was a real RNA recombination at all, or whether the crossover event happened among cDNA copies of the viral RNA.

As mentioned already in the introduction C. Biebricher and R. Luce among others developed a cell free replication systems to study molecular evolution [5],[4]. Biebricher et al. could show that during replication of the optimized RNA species MNV-11 a longer RNA sequence called SV-11 originated in the presence of higher salt concentrations [5]. Sequence analysis revealed that the SV-11 is a palindromic of MNV-11 obviously created through a recombination event between the plus and the minus strand of the virus. Biebricher explained the formation of the SV-11 product via a template switching mechanism of the $Q\beta$ replicase. If the template strand forms an unfavourable secondary structure, the replicase falls off the strand, eventually continuing on the same strand, at a different position or on a different strand. This requires complementarity between the $3'$ nucleotides of the aborted replica chain and the new template. A transesterification mechanism as discovered by Zaug and Cech [72] was excluded by incubating the MNV-11 RNA without the enzyme, which did not increase the formation of SV-11. According to these experiments, non-homologous RNA recombination can obviously take place completely without a big enzymatic apparatus, as it happens in an *in vitro* experiment. Munishkin, Voronin and Chetverin found a production of various short RNA chains by $Q\beta$ replicase [51]. However, according to Chetverin [12] $Q\beta$ recombination does not require homology. Chetverin used a reporter system of $5'$ and $3'$ fragments of an RNA species which is replicated by $Q\beta$. The fragments themselves are not replicate-able, but some of their recombinants are. He grew the RNAs on agarose containing the replicase and rNTPs, where each RNA colony on the agar represents the offspring of a beneficial recombination event. Sequencing revealed that all recombinants were non homologous. As a control the experiment was repeated in the presence of dNTPs and reverse transcriptase, an enzyme capable of template switching, which resulted only in homologous recombinants. Chetverin suggested a mechanism where the $3'$ terminus of the $5'$ fragment attacked phosphodiester bonds within the $3'$ fragment. Indeed a modification of the $3'OH$ of the $5'$ fragment inhibited the formation of recombinants.

Obviously RNA recombination can happen in a homologous and a non homologous way. Unlike DNA recombination it does not require an elaborate enzymatic system, but the simple constituents of an *in vitro* replication reaction. RNA can be recombined via a template switch during replication or independent of replication through a transesterification of existing RNA chains. Moreover a rare recombination via trans-esterification is found even when no proteins are present.

These forms of recombination could have been the ancestors of all present-time forms of cross-over. Non-homologous RNA recombination is usually not beneficial because it combines completely unrelated sequences independent of phase or homology, thereby typically disrupting functional elements. The fact that this form of recombination can be found *in vitro* but not *in vivo* can be explained by the in general negative effect of simple recombination. Living

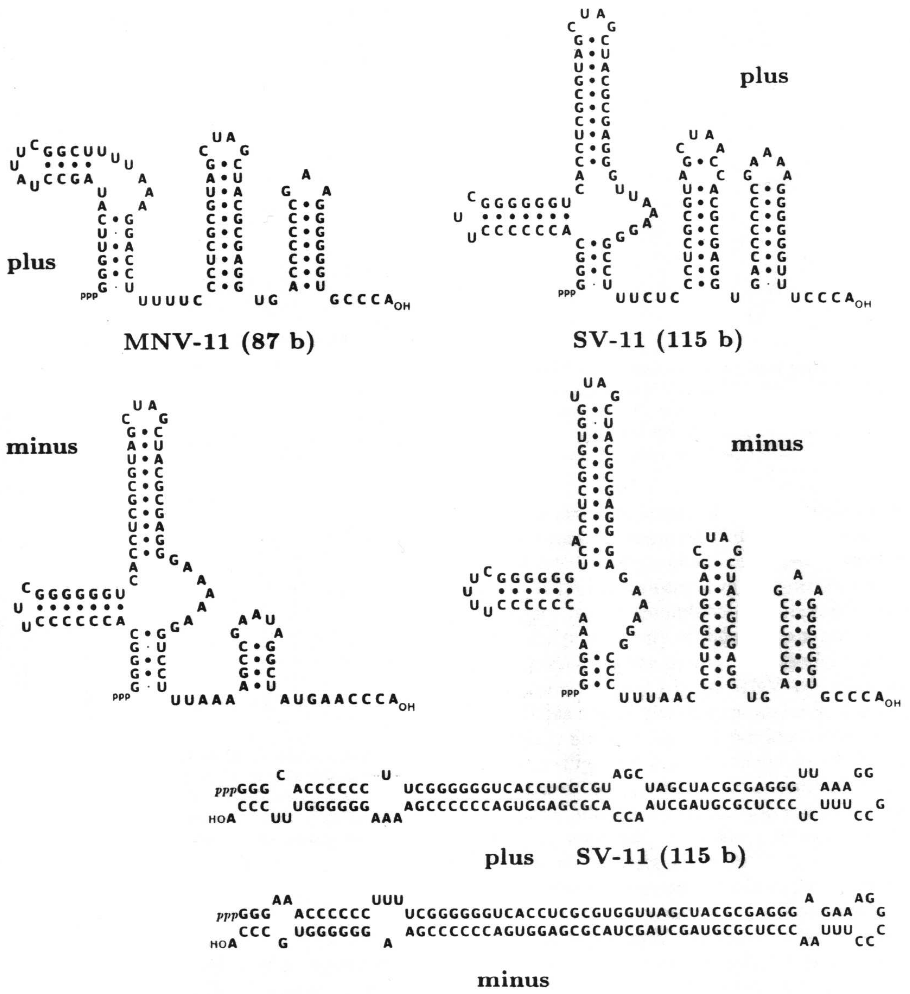


Figure 3: The secondary structures of the original template used by Biebricher, MNV11 and the recombinatorial by-product of replication SV11. The right upper graphs show the metastable structure of SV11 that is able to replicate. The lower graphs display the mfe structure of SV11, which is not recognized by the replicase any more. Figure taken from [5, Figure 5].

organisms adapted to avoid this event by inventing e.g. a more complex recombination machinery or single strand binding proteins which avoid the formation of a secondary structure causing the polymerase to fall off the strand.

2.2 Recombination Considered in a More Formal Way

Taking the doubts about the real role of recombination in evolution into account, one should ask whether it is possible to compare the efficiency of recombination as an evolutionary search strategy with that of point mutations. The publications of P. Stadler, R. Seitz and G. Wagner focus on exactly this problem [64, 68, 65].

Recombination can be regarded as a transition operator in an evolutionary optimization problem. The generic structure of an evolutionary problem can be written as

$$x' = S(x, w) \cdot T(x, t)$$

where x is the vector of haplotype frequencies and $S(x, w)$ is a term describing the selection forces acting on x. The parameters w form the fitness function, which is a mapping from the set of types into the real numbers. The second term $T(x, t)$ describes the transmission processes via the probabilities of the transformation of one type into another, by mutation or recombination. Such evolution models can be considered as dynamical systems of genotype frequencies, which live on an algebraic structure determined by the genetic processes, like mutation or recombination. However it turned out that the approach via dynamical systems theory is not suitable for more complex problems with a large number of types. In these cases the algebraic structure on which the system is realized, becomes more important than the dynamic equations themselves. E.g. is the global behaviour of an evolutionary optimization mostly dependent on the accessibility of superior genotypes via mutation or recombination of those already realized in the population. Accessibility of a particular type is mainly determined by the number of steps necessary to reach it and the fitness of the intermediates.

The process of an evolutionary optimization becomes more descriptive by the introduction of the term *landscape*. An optimization resembles the walk on a landscape, where uphill moves are preferred. A landscape consists of three ingredients: a set of genotypes V, a set of genetic operators like mutation or recombination and a fitness function. For an introduction about fitness functions and a detailed description of the fitness functions used in this work please refer to chapter 3.2. The genetic operators χ induce a topological structure on V which is called the configuration space (V, χ). A simple case of such a configuration space, is the one created by the set of strings of characters (like the nucleotides A, G, C and T) of specified length n with point mutations. This configuration space can be represented by a simple graph, where the types form the nodes and the edges connect all pairs of nodes which are separated by a single point mutation. The resulting graphs are called Hamming graphs. Walks on landscapes of hamming graphs were of some practical importance for this work and are described briefly in chapter 3.3. The ease with which Hamming graphs can be constructed comes from the simplicity of the point mutation operator, which has only one argument and gives only one result. The set of types accessible by point mutation from one particular type, defines a topological neighborhood. Unfortunately

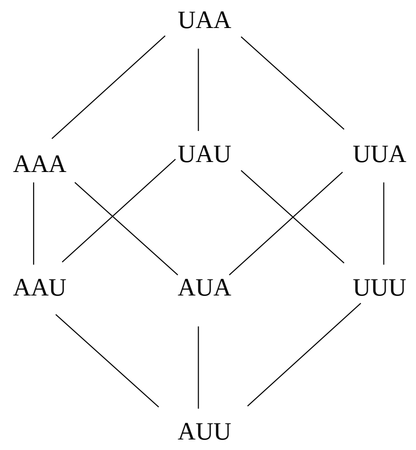


Figure 4: A simple hamming graph composed of all sequences of length 3 and the alphabet $\mathcal{A} = \{A, U\}$.

this simplicity applies not to recombination. First of all recombination is an operator with two arguments. Additionally the probability of a particular recombination event depends on the frequency of matings between the two types, which implies that it is population dependent.

The most immediate consequence of these restrictions is that the configuration space induced by recombination cannot be represented as a simple graph, with the set of genotypes representing the vertices. This can be obviated by two approaches. One can either change the nature of the vertices or of the edges. Culberson and Jones chose the first approach and substituted the types as vertices by pairs of vertices which results again in a graph since each elementary recombination event (an event at a fixed position) yields again up to two different strings. However such graphs are not comparable with hamming graphs and thereby no means for the comparison of recombination and point mutation. Gitchoff and Wagner [27] went the opposite way and could show that it is possible to represent recombination spaces in form of hypergraphs, where the vertices remain the same as in Hamming graphs, but the hyper-edges are the sets of all possible recombinants that can arise from the recombination of the types connected by the hyperedge. Whereas the latter approach is maybe not as graphic as the first, it could be shown by the latter representation that string recombination spaces and point mutation spaces are homomorphic. The only flaw about recombination hypergraphs is that they do not indicate which pair of types produces which set of recombinants. Therefore *P-structures* were invented by P. Stadler and G. Wagner [64]. P-structures are mappings from pairs of types to the edges in the hypergraph. Using this approach it was possible to apply Fourier decomposition to the recombination spaces, which has already been done for point mutation spaces. With this procedure it was possible to compare the properties of recombination and point mutation landscapes.

P-Structures: Assuming a genome with n loci and each locus k has α_k alleles, the set of $\Pi_k \alpha_k$ possible genotypes shall be denoted by V. For each locus k the alleles are labelled with a letter from the alphabet $\mathcal{A}_k = \{0, ..., \alpha_k - 1\}$. A particular genotype (or sequence) $x\ inV$ can be regarded as a vector with components $x_k \in \mathcal{A}_k$.

A crossover operator is a map $\chi : V \times V \to V \times V$ with the properties: Suppose $\chi(x, z) = (u, v)$. Then for each k either $y_k = u_k \wedge z_k = v_k$ or $z_k = u_k \wedge y_k = v_k$. By abuse of notation we write $x \in \chi(y, z)$ if $x = u$ or $x = v$, i.e. if x is an offspring of (x, y). An immediate consequence of this definition of the crossover operator χ is that $\chi(x, x) = (x, x)$.

If V is a finite set with the power set $\mathcal{P}(V)$, then a *P-structure* is a pair $(V, \mathcal{R})$ where $\mathcal{R} : V \times V \to \mathcal{P}(V)$. The P-structure is called symmetric if $\mathcal{R}(x, y) = \mathcal{R}(y, x)$ for all $x, y \in V$. In a weighted P-structure a positive weight

$\mathbf{H}_{x,(y,z)}$ is attached to each triple (x, y, z) for which $x \in \mathcal{R}(y, z)$, $\mathbf{H}_{x,(y,z)} = 0$ if $x \notin \mathcal{R}(y, z)$. $\mathbf{H}$ is called the incidence matrix of the P-structure. There is a P-structure associated in a natural way with each cross-over operator χ:

$$\begin{aligned} \mathcal{R}^{\chi}(y, z) &= \{x \in V | x \in \chi(y, z)\} \\ \mathbf{H}^{\chi}_{x,(y,z)} &= \begin{cases} 2 & \text{if} \quad x = y = z \\ 1 & \text{if} \quad x \in \chi(y, z) \qquad \text{and } y \neq z \\ 0 & \qquad \text{otherwise}\,. \end{cases} \end{aligned} \tag{1}$$

Obviously $\mathbf{H}^{\chi}_{x,(y,z)} > 0$ if and only if x is an offspring of (y, z). Setting the diagonal elements $\mathbf{H}^{\chi}_{x,(x,x)} = 2$ implies immediately that $\sum_x \mathbf{H}_{x,(y,z)} = 2$, since any crossover operator produces exactly 2 offsprings from a pair of parents. If $y = z$, the offspring is simply counted twice.

Cross-over Operators The only type of crossover operators which shall be considered here is defined on the set $V = \mathcal{Q}^n_\alpha$ of strings with length n and over an alphabet consisting of α letters. The mapping $\times_k : V \times V \to V \times V$ is called an elementary operator if:

$$\times_k = ((y_1, y_2, ..., y_{k-1}, x_k, ..., x_n), (x_1, x_2, ..., x_{k-1}, y_k, ..., y_n))$$

for $1 \leq k \leq n$. $\times_k$ describes the action of crossover at point k, in particular $\times_1$ is the identity mapping. A second useful basis of crossover operators is the following:

$$\Diamond_k = ((x_1, x_2, ..., x_{k-1}, y_k, x_{k+1}, ..., x_n), (y_1, y_2, ..., y_{k-1}, x_k, y_{k+1}, ..., y_n))$$

$1 \leq k \leq n$, which swap only the position k between x and y. Obviously we can write $\times_k = \Diamond_1 \circ \Diamond_2 \circ ... \circ \Diamond_{k-1}$ and $\Diamond_k = \times_k \circ \times_{k+1}$. The basic algebraic properties of the single position operators $\Diamond_k$ follow directly from the definition: they commute, because different positions do not interfere with each other, $\Diamond_k \circ \Diamond_l = \Diamond_l \circ \Diamond_k$, moreover they are all involutions, i.e. $\Diamond_k \circ \Diamond_k = \times_1$. Any general crossover operator χ is a finite, but arbitrary composition of elementary recombination operators.

Recombination Operators A *recombination operator* is a family $\mathcal{F}$ of cross-over operators that act on $V \times V$ wit the probability $\pi(\chi)$. The weighted

P-structure associated with a recombination operator $\mathcal{F}$ is then:

$$\begin{aligned} \mathbf{H} &= \sum_{\chi \in \mathcal{F}} \pi(\chi) \mathbf{H}^{\chi} \\ \mathcal{R}(y,z) &= \bigcup_{\chi \in \mathcal{F}} \mathcal{R}^{\chi}(y,z) = \left\{ x \in V \middle| \exists \chi \in \mathcal{F} \,:\, x \in \chi(y,z) \right\} \end{aligned} \tag{2}$$

If for χ holds $\chi \in \mathcal{F}(\mathcal{R})$ it is said that χ contributes to $\mathcal{R}$.

Recombination Structures According to Gitchoff and Wagner [27] a P-structure $(V, \mathcal{R})$ is a *recombination structure* if for all $x, y, z \in V$ the following axioms hold:

1. $\mathcal{R}(x,x) = \{x\}$.
2. $\mathcal{R}(x,y) = \mathcal{R}(y,x)$.
3. $\{x,y\} \subseteq \mathcal{R}(x,y)$.
4. For all $z \in \mathcal{R}(x,y)$ holds $|\mathcal{R}(x,z)| \leq |\mathcal{R}(x,y)|$.

Considering the previous two types of elementary cross-over operators, four different recombination operators seem possible. They all fulfill the axioms for recombination structures.

$\mathcal{R}_1$ *One-Point Crossover* is defined by the collection of all elementary operators, $\mathcal{F} = \{\times_1, \times_2, ..., \times_n\}$.

$\mathcal{R}_2$ *Two-Point Crossover* consists of all compositions $\times_k \circ \times_l$, $k, l \neq 1$. For technical reasons the identity is included as well.

$\mathcal{R}_\Omega$ *Uniform Crossover* allows for all possible recombinations to take place, i.e., $\mathcal{F}\left(\mathcal{R}_\Omega\right) = \left\{ \xi = \prod_{j \in J} \Diamond_j | J \subseteq \{1, 2, ..., n-1\} \right\}$.

The only cross-over operators utilized in this work are one-point crossover operators. They are only chosen for the sake of simplicity. In natural systems two-point and multipoint crossovers play an important role. E.g. is the consideration of two-point and multi-point crossover important for the creation of linkage maps in genetics. Two-point and multi-point crossover obscure the recombination frequencies and lead to an underestimation of distances. Usually this problem is alliviated by measuring linkage between genes close to each other, where the probability of two-point crossovers is very small or by using so called *tetrade analysis*[28].

An example of uniform crossover is the "trivial" crossover occuring in sexually reproducing diploid organisms during meiosis, i.e. the fair transmission of chromosomes into the gametes. This form of crossover is not considered any further in the course of this work, but gives an upper bound for the maximal recombination freuquency between two distant loci. The maximal recombination frequency between two loci is 50%, which is the probability to transmit a maternal and paternal allele instead of two maternal ones or two paternal ones into the gamete.

2.3 Genes in Pieces – Recombination and Introns

Though it had been predicted for years that the human chromosome is too big to encode only the estimated number of genes the outcome of the first results of the human genome project was astonishing. Only 5-10 % of the human DNA are transcribed and in consequence translated into proteins! In 1977 several groups found out that genes were not only separated by large parts of nonsense DNA, but that the genes themselves were interrupted by sequences which were missing in mRNA and protein [3, 6, 7, 13]. Following a suggestion by Walter Gilbert [23], these interrupts were called *introns*, for "intragenic regions". The information bearing parts resulting in the amino acid sequence of the proteins were called *exons*, as a shortcut for expressed regions. The absence of the intronic sequences in the observed mRNAs gave rise to the question of the mechanism which caused the disposal of the sequence stretches. Francis Crick proposed four possible pathways [14].

1. A DNA rearrangement displaces or eliminates the intronic sequences. The DNA in the germ line should remain unaltered.

2. The DNA remains unaltered, but the RNA polymerase skips the introns, thereby producing a primary transcript consisting of exons only.

3. Each exon is transcribed separately and the separate pieces of RNA are ligated together.

4. The RNA polymerase produces a transcript of the whole gene, containing exons and introns. The introns are removed in a processing step prior to translation. This RNA processing is called splicing.

The first case was indeed found in one system. The two stretches of DNA coding for the light chain of the immunoglobulin in the mouse are found to be wide apart in germ line cells. In somatic cells producing the protein the two exons lie much closer. However the immune system is a special case and this at any rate interesting fact shall not be discussed any further. The second and the third proposed mechanisms have not been discovered to be used in any biological system so far. The last mechanism is the most important one, which is used among all intron bearing species.

Introns can be divided into four groups, all of which work in the way of Crick's fourth proposal. The evolutionary ancient introns work independent of any enzymatic splicing mechanism and are therefore called self-splicing introns. Dependent on their splicing mechanism they form at least three groups. Today's most prominent intron type are the spliceosomal introns, which are evolutionally related with group II self-splicing introns [10]. Spliceosomal introns need a complex of small nuclear ribonucleoproteins (snRNPs) for the splicing process. Today it is known that the splicing process of spliceosomal introns is a two step enzymatic reaction, consuming energy in form of ATP. The high fidelity of the splicing process is maintained by a complex of snRNPs, which brings the 5' end of the intron close to a nucleotide in the 3' terminal region of the intron. The nucleotide attacks the 5' end of the intron and cleaves it. In a second step the 3' terminal OH of the first exon displaces the intron and the two exons are joined. The 5' end of the intron remains at the attacking nucleotide, giving the intron a lariat like shape, which is visible under the electron microscope.

The determination of complete genomes of many eubacteria and the progress in sequencing the genomes of multicellular animals made clear that introns are not present in all organisms [54]. Whereas mammals have a high intron to exon ratio and a low coding density, most eubacteria and archea are nearly devoid of any intergenic sequences. About 90% of the genomes of eubacteria and archea are dedicated to protein coding genes and they have a coding density of about 1000 genes per Mbp. The simplest eucaryotic organism, the unicellular *Saccharomyces cerevisiae* seems lavishly compared to the procaryotes. Open reading frames occupy only 70% of the genome, the coding density drops to 400 genes per Mbp. The genome of the nematode *Caenorhabditis elegans*, which has about 6 times the size of a yeast genome is an example of a well studied genome of a simple multicellular organism. The coding sequence accounts for 26% of the genome and the coding density is 200 genes per Mbp. Of special interest is the genome of the pufferfish *Fugu rubripes*, because it has roughly the same number of genes as the human. The *Fugu* genome has the same intron/exon organization as mammals, but the introns are much shorter than their mammalian counterparts. The coding density and the fraction of coding sequence are more or less the same as in *C. elegans*. The fact that the genome of *Fugu* is about 7.5 times more compact than the human genome allows to estimate the coding density of the human genome which is about 20 genes per Mbp and a fraction of coding sequence in the genome of less than 5%. A comparison of the intron frequencies in various species suggests that the fraction of introns in the genome correlates well with the genome size. Additionally not only the amount, but also the types of introns are distributed unequally among various phyllogenetic lineages. Group I introns were originally restricted to tRNA genes, but are meanwhile found in several genes and are phylogenetically widespread [10]. They were found in eubacteria, bacteriophages, eucaryotic organelles and nuclei [39, 59]. Group II introns are much more restricted in their distribution, they were found in cpDNA (chloroplast) and higher plant mtDNA (mitochondrial), and a minority of introns in fungal and algal mtDNAs. Recently, they were also found in *Proteobacteria* and *Cyanobacteria* [48, 20, 36]. Spliceosomal introns are the most widespread class of introns. They are found in most eukaryotic genomes, but are missing in some of the most primitive eucaryotic representatives, the protists [43]. Procaryotes are completely devoid of spliceosomal introns.

The discovery of introns and their unequal distribution, especially between unicellular and multicellular organisms, gave rise to the question where introns came from and whether they had any biological purpose. Three early ideas about the evolution of introns have been reviewed by Francis Crick [14]. All models assumed for simplicity that the splicing signals lie in the terminal regions of the intron, which later on proved to be true. The first model assumes that the first intron came into existance in a gene which had been already transcribed. The splicing signals would have been created by random mutation and an already existing splicing machinery would have recognized these signals. Alternatively the first intron could have been created by already existing signals in the DNA which were suddenly recognized by a newly created splicing enzyme. The second mechanism explains the existence of introns, by the assumption of a DNA insertion process which creates signals closely related to splicing signals. A few subsequent mutations could have been sufficient to initiate splicing. The third proposed mechanism assumes that new introns are created by translocating an exon with its flanking introns into an already existing intron, thus producing two introns where there was only one before.

Which of the proposed mechanisms is the true one, or whether the origin of introns involves all three proposed mechanisms, has not been decided yet. The important difference between the proposed mechanisms is, that the first two do not predict a special time point for the invention of introns. The third mechanism in contrast is based on the assumption that introns existed prior to the construction of more complex proteins. Todays controversy about the origin of introns is mainly fought among the supporters of the introns late hypothesis which is the synopsis of the first two mechanisms and the protagonists of the introns early theory, which is based on the third proposed process.

Introns Early Hypothesis or the Exon Theory of Genes Based on the third suggested mechanism of intron development, Darnell [15], Doolittle [19] and later on Walter Gilbert [23] created the introns early hypothesis. This explanation for the origin of introns assumes that (spliceosomal) introns pre-date the divergence of procaryotes and eucaryotes. Walter Gilbert extended this hypothesis to a model about the origin of life [24]. According to Gilbert primordial forms of life were autocatalytic RNA molecules which acted as a holder of information and as enzymes catalyzing their own reproduction simultaneously. Later on transfer RNAs evolved, which allowed the catalytic construction of polypeptides, resulting in first proteinaceous enzymes which could have taken over the enzymatic properties of the RNA molecules. Finally the invention of DNA displaced RNA as the information keeper, leaving RNA only in its intermediate role that we know today. This concept of primordial evolution is important for the theory of introns, because Gilbert concluded that the intron/exon structures of genes were the leftovers of this RNA world. The parts of autocatalytic RNAs which were necessary to control the specific condensation of amino acids, became todays exons, whereas the rest formed the ur-introns. The enzymatic activities of self splicing introns are taken as an argument for their relationship with ancient autocatalytic RNAs. Gilbert extended this model even further to explain the evolution of more complex proteins. Obviously the ur-exons were functionally self-contained, autonomously folding elements, otherwise they would not have survived. Gilbert claimed that the ur-exons should have worked as building blocks for more complex proteins [25]. By means like intronic recombination, transposition or retro-transposition the exons separated by introns could be brought into new combinations. This process of recombining exons to form proteins with new functionalities is called exon shuffling. Such a mechanism of recycling already approved modules could have greatly sped up evolution, because the negative effects of pure random mutation are avoided.

The introns early hypothesis (IE) provides a nice explanation how complex multidomain proteins have evolved. However those proteins are most probably relatively new. So how can one prove that introns are as old as life, or at least older than the last common ancestor (LCA) of procaryotes and eucaryotes? The supporters of this hypotheses base their arguments mainly on the position of introns in genomes. Ancient introns should separate building blocks of proteins, which are functionally and in terms of folding efficiency self-contained. So one would expect to find ancient introns preferentially at positions in between such building blocks and not amidst them. Gilbert and co-workers examined the positions of the introns in a very old gene, the triosephosphate isomerase (TPI) gene. TPI is part of the glycolytic pathway, which is common among all beings. The TPI gene shows introns at identical positions in plant and animal genes which would suggest, that these introns arose before the separation of plants and animals [46, 67]. Based on the modules and the intron/exon structure of TPI Gilbert made the prediction that an additional intron should be found at a position breaking up a TPI exon not apparently representing an individual protein module. When such an intron was found in the mosquito *Culex*, it was taken as a strong evidence for the introns early hypothesis. However further analysis of TPI genes and protein structures were conflicting in their results and interpretations [66]. A recent discovery of new introns in TPI questions its role as a strong argument for the exon theory of genes [44]. Analysis of other different genes revealed that introns do not always separate α-helices and β-sheets as one would expect for ancient introns [66]. Introns-early supporters replied that probably the building blocks of proteins are different from their secondary structure elements. A possible identification of modules in proteins is to search for regions of the amino acid chain which fold into a sphere of defined diameter, an idea which was developed by Mititko Go. De Souza and Gilbert [17] compiled so called Go plots for proteins by calculating the pairwise distances of all residues in a protein structure. If the pairwise distances are written down in a $n \times n$ matrix and n is the number of residues in the protein, modules can be identified as blocks of distance values below a certain threshold. De Souza and Gilbert chose 28 Å and 33 Å as module diameters. Because the modules tended to overlap, the overlapping regions were called linker regions. The introns-early approach would suggest, that introns lie preferentially within those linker regions, which was indeed found by de Souza and Gilbert. A repetition of the experiment with arbitrary module diameters [18] revealed peaks of highest significance for the correlation at diameters of 21, 28 and 33 Å. The experimenters concluded from these results that the ur-exons had a typical length of 15 (21Å), 22 (28Å) or 30 (33Å) amino acids. The second main argument for the introns early hypothesis is the phase distribution of present time introns. Introns of protein genes can be classified according to their position relative to the reading frame of genes [60]. (i) introns in the 5' or 3' non-coding regions of genes; (ii) introns lying between two codons, called phase 0 introns; (iii) introns lying between the first and the second base of a codon (phase 1 intron); (iv) introns lying between the second and the third nucleotide of a codon, called phase two introns. If introns were inserted into previously uninterrupted genes, as the introns late view advocates, the structure of the gene product should not be affected by the insertion. Therefore the chance for an intron to survive should be the same independent of its phase and introns should be distributed equally among genes. The introns-early view predicts a nonrandom phase distribution for introns, because an intronic recombination event among introns of different phase will cause a frame-shift mutation, which is polar and only rarely beneficial. This does not necessarily mean that phase 0 introns are privileged ones, because symmetric exon shuffling among phase 1 or 2 introns differs from phase 0 only in the creation of at most one point mutation in the amino acid chain. Additionally one would expect to find primarily symmetric exons, i.e. exons which lie between introns of the same phase, which should make exon shuffling easier. Consequently, if late insertion of introns is true one would expect to detect

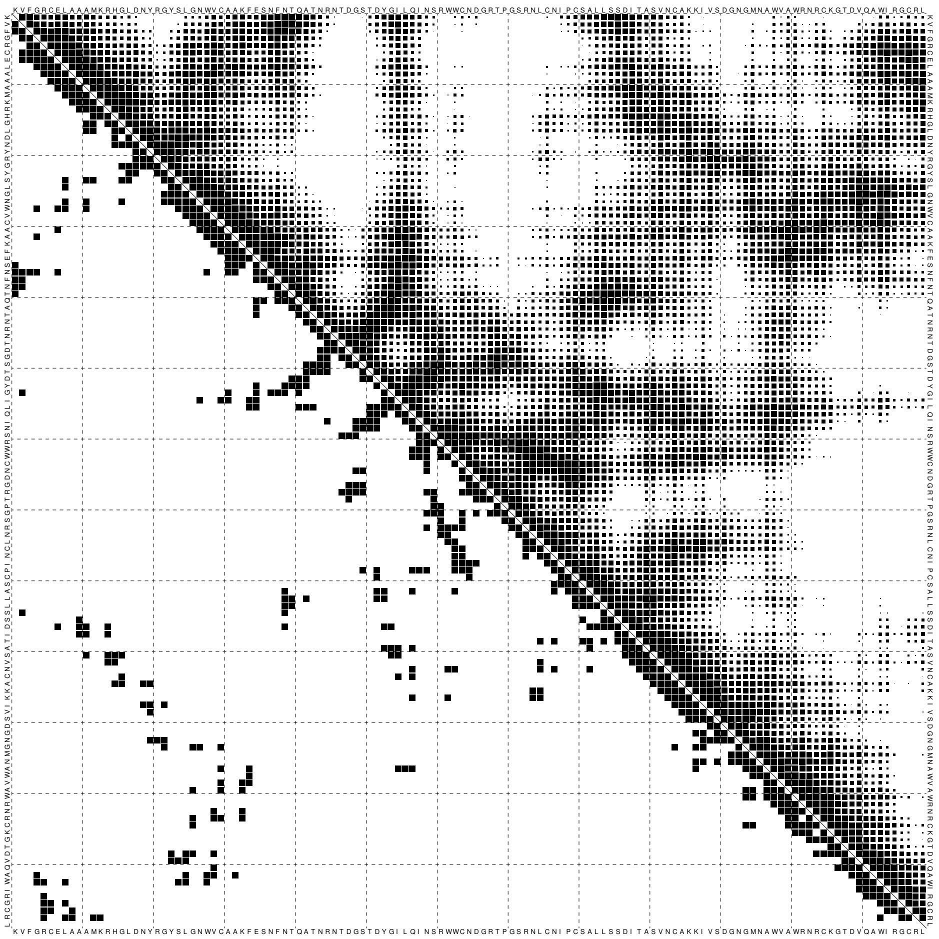


Figure 5: Go-plot and conatct plot of 1LSE. The upper triangle of the dot plot displays the Go-plot of the structure. The area of the squares, A corresponds to the distance $dist$ between the squares: $A = 1 - dist/ct$, the cutoff ct is set to 23Å. The lower triangle of the dotplot shows contacts between residues. Black squares indicate that the pair of residues has been identified as a contact by the tesselation procedure (cf. section 3.2.2).

each class of introns with probability 1/3 and each class of exons with probability 1/9. Long *et al.* analyzed a database of 296 genes containing 1496 introns and 1200 internal exons [45]. He obtained the database by purging an originally bigger one for homologues greater 20% and eliminated all introns and exons which did not lie within an ancient conserved region (ACR). ACRs are regions of eucaryotic genes which match their procaryotic paralogs with high score. Long found highly unequal proportions of the three intron classes: 48% phase zero, 30% phase one and 22% phase two. The observed intron associations, showed that symmetric exons are much more frequent than would be expected for random distribution. Interestingly not the (0,0) exons showed the highest excess over the expected value, but the (1,1) exons were the most frequent ones. The result of the phase correlation statistics is maybe the strongest argument for the antiquity of a fraction of introns.

The Introns Late Hypothesis (IL) is a synthesis of the first two proposals discussed on page 19. It is the exact opposite of the introns early hypothesis and assumes that introns originated relatively late in evolution by random insertion. Advocates of this hypothesis can be subdivided in a group which rejects the idea of early introns and the assembly of all genes by exon shuffling, but believes in the shuffling origin of some newer proteins; and a group, which turns down both, the early origin of introns and any exon shuffling. The hypothesis rests mainly on the phyllogenetic distribution of introns, which cannot be properly explained by the introns early approach. Supporters of this hypothesis conclude from the fact that spliceosomal introns are completely missing in procaryotes and some primitive eucaryotes, that introns were invented after the separation of the procaryotic and the eucaryotic lineage. A simple explanation given for the absence of spliceosomal introns in bacteria that splicing requires the separation of the unspliced RNA from the peptide synthesizing ribosomes, which is not given in bacteria. This does not proof to be true, because such introns were found in mitochondria, which do not have a nucleus either [14].

On the contrary, the IE hypothesis has some difficulties with these facts. It assumes that introns in procaryotes got subsequently lost, caused by the need for shorter reproduction periods, which implies faster DNA synthesis. Procaryotes unlike eucaryotes have only one origin of replication, which means that the time necessary for the duplication of DNA is directly proportional to the size of the genome. A possible mechanism for the loss of introns is retrotransposition, which is the transcription of a separated gene into pre mRNA, and a reverse transcription after the splicing process, yielding intron free DNA, which can integrate into the genome via recombination. Proponents of the IL approach question this model of intron loss, because no remnants of a splicing machine have ever been found in procaryotes.

Cavalier-Smith developed a model for the evolution of spliceosomal introns in eucaryotes based on the IL approach, the *seed hypothesis* [10]. The splicing mechanisms of group II introns and spliceosomal introns are very similar, which suggests their evolutionary relation. This was confirmed by a recent experiment: the deletion of a stem loop in a group II intron was complemented by the addition of U5 snRNA, which is a part of the spliceosomal machinery [30]. Based on these similarities, Cavalier-Smith argued that a group II intron might have been donated to a eucaryotic nucleus via lateral transfer. A subsequent fragmentation of the intron ended up in what is considered a spliceosomal intron today. This model was backed by the lack of spliceosomal introns in the earliest branches of eucaryotic evolution, the *Archezoa*, which do not have any organelles. Meanwhile, however, it could be shown that *Archezoa* have mitochondrial genes in the nucleus. Obviously those protists contained mitochondria, which were subsequently lost. This and the fact that no group II intron, which could have been the ancestor of spliceosomal introns, has been found yet, leaves a lot of doubts about the mitochondrial seed hypothesis.

One of the main successes of the IE hypothesis was the prediction of an additional intron in the TPI gene and the finding of this intron in the genome of the mosquito *Culex*. Supporters of the IL hypothesis turned the *Culex* intron into an argument against IE and claim that this intron is not at all an ancient intron, but is relatively new. This view is founded by the distribution of the intron in the relatives of *Culex*. The intron is only present in *Aedes*, which is a very close relative of *Culex*, but not in the looser relative *Anopheles* or in any other sequenced insect TPI gene [44, 38]. The defense of the IE is once more that the ancient intron got lost in the other mosquito breeds. Parsimony, however, contradicts this defense.

Today it is widely accepted that introns undergo changes continuously, are inserted and lost. Even supporters of the IE approach admit that most of the present time introns are of newer origin, or have changed so much that they cannot be properly identified as ancient introns [16, 57]. Most supporters of the IL do not agree that parts of the present time introns could be of newer origin and parts ancient, but persist that introns developed late in evolution. Ongoing findings of correlation between the intron/exon structure of genes and the secondary structures of the corresponding protein structures seem to confirm the introns early view [9, 35]. It seems to be a paradoxon of genomics that the phyllogenetic distribution precludes ancient introns, but a pure introns late view cannot explain the correlation with protein structure. If our view of recombination proofs to be true, it could help to solve this paradoxon. We expect to find introns at module boundries, but allow the dynamic insertion and deletion of introns. Our model does not require that the introns have the same age as the genes they interrupt.

Exon shuffling and Recombination Exon shuffling is a model invented by Gilbert [25] which explains the evolution of more complex proteins from simple polypeptides. Gilbert assumed that new proteins were created by intronic recombination among the ur-introns which separated the rather short ur-exons. According to Gilbert this should have happend at the transition from the RNA world to the protein world. However if this imagination is true, which kind of introns and which kind of recombination are involved in this process? It is almost sure that spliceosomal introns originated relatively late, most probably after the separation of the procaryotic and eucaryotic lineages, from group II introns [10]. Thus, the only known introns which could have played a role in the RNA world are self splicing introns, presumeably very close relatives of todays group I or group II introns. Self splicing introns however are not very robust in their nucleotide sequence, the splicing activity can be destroyed rather easily [54]. As we pointed out earlier, non-homologous recombination events are usually disadvantageous, implying that recombination between non-homologous self splicing introns, which are dependent on the formation of certain secondary structure, will destroy the splicing activity in the majority of cases. Consequently the only form of recombination that is appropriate for primordial exon shuffling is homologous recombination. Homologous recombination, however, seems to play only a small role among RNA molecules. A restriction of recombination events to homologous ones, requires an elaborate enzymatic apparatus (cf. 2.1.1) and was certainly not available at the transition form the RNA to the protein world. We propose an introns intermediate view: introns were inserted after the invention of the spliceosome. The bias in the position distribution of introns could be due to the regulatory effect of introns on the impact of non-homologous recombination. We agree with Patthy that after this process the invention of exon shuffling via homologous or non-homologous intronic recombination is responsible for the creation of many metazoan proteins [54].

3 Methods

The effect of recombination on the modularisation of genes is studied on the level of RNA genomes and on the level of protein encoding DNA genomes. In both cases we consider a population of haploid species. In the case of RNA the species have an RNA genome, bearing only one gene, which is an RNA sequence with a certain biologic activity, represented by its minimum free energy secondary structure. The initial population consists of a number of species with unrelated (random) genotypes but the same phenotype, i.e. the same minimum free energy structure. The species are allowed to produce offspring via recombination and the phenotypes of the offspring are compared with the initial phenotype. The effect of recombination is then evaluated by counting the number of species retaining the original phenotype and by inventing a distance measure between the phenotypes to calculate the mean distance between the phenotypes of the parents and the offspring. . In the protein experiments the species have a DNA genome encoding one protein. Again the biological activity is determined via the structure of the biopolymer encoded by the gene. In contrast to the RNA experiment it is not possible to make an exact structure prediction – neither secondary nor tertiary – for amino acid sequences, so the considered phenotype is not the structure itself, but the ability of an amino acid sequence to fold into a given spatial protein structure. An analogous experiemnt as described for RNA species is performed with the protein species. A population of species with unrelated genotype but with the same ability to fold in the target structure reproduces via recombination and the phenotypes of the offspring are compaired with the initial phenotype. As the phenotype is represented by a real number the invention of a distance measure is trivial.

The modularisation of protein genes under recombination is tested in an evolutionary simulation. A population of protein species is kept growing in a flow reactor like environment. Species reproduce with different rates dependent on their fitness, which is their phenotype. Whenever a replication takes place, point mutation, recombination and the insertion and deletion of introns can take place. The evolutionary optimization and the placement of the introns averaged over the whole population are monitored during the simulation.

3.1 RNA Secondary Structure Prediction

Usually RNA molecules do not form double strands like DNA but rather fold back on themselves yielding a spatial structure. This spatial structure or tertiary structure defines the biological function of the RNA molecule. Unfortunately the exact tertiary structure depends mainly on environmental conditions and because of its many degrees of freedom cannot be predicted by computational means. Prior to the formation of the tertiary structure the RNA molecule forms a pattern of complementary base pairings which is called the secondary structure. The secondary structure is believed to provide a scaffold of distance constraints, which guide the formation of the tertiary structure. Therefore biological function and secondary structure are tighly related as well. What makes the secondary structures of various RNA sequences so interesting for this project is that they can be calculated easily via a dynamic programming algorithm [74]. RNA sequences are compatible with many secondary structures which means that they can form various structures. For our purpose we will assume that the biologically active secondary structure of a sequence is the most stable structure the sequence can form which is the structure of minimal free energy, called mfe stucture in the following. The comparison of the mfe structures of related sequeneces (e.g. related via a recombination event) gives us a tool to test the effect of genetic operations like recombination or point mutation on an artificial biological function. Because of its importance for the RNA recombination statistics described in chapter 4.1.1, the algorithm for RNA secondary structure prediction shall be discussed briefly.

A secondary structure $\mathcal{S}$ is defined as the set of base pairs (i, j), with $i < j$, such that for any two basepairs (i, j) and (k, l) two conditions hold: $i = k$ if and only if $j = l$ and $i < k < l < j$ or $k < i < j < l$. The first condition simply means that each base can particpate in at most one base pair. The second condition is called *non pseudoknot condition* and is required to allow the solution of the problem via a dynamic programming algorithm. Pseudoknots are base pairs between self-contained secondary structure elements, e.g. two otherwise separated loops. The first dynamic programming algorithm to solve the prediction problem was the maximum matching algorithm developed by Ruth Nussinov [53, 52], which yields the structure with the highest number of base pairs. However these structures are clearly not equal with the mfe strcutures because the H-bond itself does not stabilize the structure, it does not matter whether it is formed with the complementary base or with a water molecule. The main stabilizing force is the stacking effect. The aromatic systems of the bases assemble parallel to each other to maximize the overlap of the delocalized π-electrons. For the calculation of the minimum free energy the secondary structure is dissected into loops and calculated as the sum the energy contributions of all loops. The only stabilizing loops are stacks, which are considered as loops of grade two and size one and bulges, all bigger loops like hairpin loops or multiloops act destabilizing. Figure 6 shows an examplary loop decomposition.

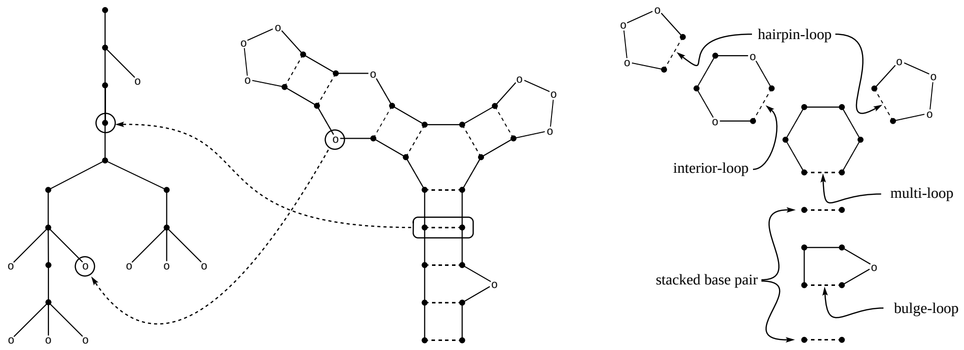


Figure 6: Loop decomposition of an RNA secondary structure. The representation of RNA as a planar graph is the most common one (middle). The tree representation of the same secondary structure as in the planar graph (l.h.s). The loop decomposition of the secondary structure graph in the middle (r.h.s.). The closing basepairs of variaous loops are indicated by dotted lines.

All computations of RNA secondary structure and sequences by inverse folding were performed with algorithms from the *Vienna RNA Package*, [32, 74], the current set of parameters has been taken from [47].

3.2 Fitness Functions for Proteins

In course of this work repeatedly a crucial decision has to be made: whether a particular sequence which usually originated from an other sequence by performing editing operations like point mutations, insertions etc is still capable of a biological function. We will call a measure for this "ability" the fitness of the sequence. So what we are looking for is a function $F : \mathcal{A}^n \longrightarrow \mathbb{R}$, with $\mathcal{A}$ being the alphabet and n denoting the sequence length.

Actually not the ability of the sequence to perform any biological process is of interest, but a specific biological activity. One ussually assumes that the function of a biopolymer is defined by its spatial structure and consequently the ability of a sequence to fold into a given structure defines a fitness function $F : (S, \psi) \longrightarrow \mathbb{R}$, with the sequence $S \in \mathcal{A}^n$ and ψ denoting a fold.

Following Anfinsens's pinoneer experiments in 1973 [2] the so called folding postulate states that "in equilibrium the native state of a protein-solvent system corresponds to the global minimum of free energy". In concequence of this postulate one could compute the fitness of a sequence by comparing the free energy of the native sequence and its fold with the free energy of the test sequence and the same fold.

Of several possibilities to calculate free energies of molecules ab-initio and semiempirical methods drop out because of their computational effort which makes them only applicable to small molecules. In principle free energies could be calculated with molecular mechanics force fields, though they are not "cheap" as well, especially if the solvent is taken into count, which one should bescause the major part of stabilizing energies of a protein fold comes from interactions between the polymer and its environment. Their major disadvantage is that those potentials are too fine grained for our purpose, because most of them calculate in atomic resolution. What is left are so called Knowledge Based Potentials which shall be discussed in detail in the following section.

3.2.1 Knowledge Based Potentials

Knowledge based potentials describe the energy of molecule as the sum over the energies of all residue contacts. The energy of single contact depends on its likelihood which is extracted from a database of known structures. A theoretical bases for this procedure comes from statistical mechanics. If a protein is in equilibrium state, i.e. in its energetical minimum, low energy elements must occur more frequently in 3d-structures of globular proteins than others. This relationship between the frequency of a state and its energy is described by Boltzmann's law:

$$f_{occ} \sim \exp(-E/RT)$$

T is the conformational temperature and R the gas constant. Considering this realtionship an estimate of the frequency of occurrence can be used to assign a putative energy to a sequence in a certain fold. This interpretation of knowledge based potentials was introduced by Manfred Sippl [61].

According to the relationship above one can denote the probability to find physical system in equlibrium in a particular state x by:

$$\mathrm{prob}(x) = \frac{1}{Z}\exp\left[-\frac{E(x)}{kT}\right]$$

where k is Boltzmann's constant, T the absolute temperature and Z the partition function:

$$Z = \sum_{i=1}^{n} \exp\left[-\frac{E(i)}{kT}\right]$$

If the energies of all states were known, the probability density could be calculated, or analogous the energy of a state could be computed if the probability density was accessible.

$$E(x) = -kT\ln\left[f\left(x\right)\right] - kT\ln Z$$

Whereas it is possible to obtain the frequency of occurence and thereby the probability of a state, it is impossible to compute the partition function Z, which means that an additive constant remains unknown. By extracting the distribution of probabilities from a database, a pontential of mean force of interaction can be obtained. The net potential of a contact γ can be computed, with $E(x)$ denoting the reference state of the system (the averaged energy) by:

$$\Delta E_\gamma(x) = E_\gamma(x) - E(x)$$

and expanded:

$$\Delta E_\gamma(x) = -kT\ln\left[\frac{\mathrm{prob}_\gamma(x)}{\mathrm{prob}(x)}\right] - kT\ln\frac{Z_\gamma}{Z}$$

Z_γ and Z do not depend on the state x so one can assume $Z_\gamma \simeq Z$ and therefore $-kT\ln\frac{Z_\gamma}{Z} \sim 0$. T is constrained to the temerature of the NMR or X-ray measurement of the data.

$$\Delta E_\gamma(x) = -kT\ln\left[\frac{\mathrm{prob}_\gamma(x)}{\mathrm{prob}(x)}\right]$$

After extracting parameters for all occuring contacts, a summation over all contacts yields the energy of a sequence S for a fold ψ:

$$E(S,\psi) = \sum_{\gamma} E_{\gamma}(x)$$

3.2.2 The Extended Tropsha Potential

A subgroup of knowledge based potentials are the so called Contact Potentials which measure the overall energy of a system as a sum of nearest neighbour contacts. Two promiment Members of this group are the rather coarse grained Crippen's Simplified Potential and Tropsha's Four Point Potential.

A crucial point in designing an efficient potential is the selection of valid contacts. Many potentials like the Crippen!!! potential use a heuristic approach in defining a threshold distance below which a pair of elements is called a contact. A. Tropsha avoided the arbitrariness of such a binned distance by introducing an exact!! approach from computational geometry. For this approach the protein is considered as a set of points in $\mathbb{R}^3$, for simplification usually only the C^{α} or C^{β} atoms are considered. The set of points is tesselated using the delauney triangulation resulting in a partitioning of the space included by the set into irregular tetrahedra with the points as vertices. The quadruple of aminoeacids represented by one tetrahedra are considered to be nearest neighbours.
If one counts the occurence of all possible neighborhood combinations of the aminoacids in a structural dataset, a log likelyhood function can be created. The likelihood of finding a distinct quadruple in the set of points of a protein structure can be denoted as:

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}}$$

where i, j, k, l denote four aminoacids, f_{ijkl} is the obserevd normalized frequency of occurence of a given quadruple and p_{ijkl} is the a priori expected frequency of occurence of a given quadruple. The observed normalized frequency of occurence f_{ijkl} is calculated by division of the counted occurence of a quadruple through the total number of observed quadruples and

$$p_{ijkl} = \frac{4!}{\prod_a^M t_a!} a_i a_j a_k a_l$$

where a_x is the observed number of occurence of a distinct aminoacid type divided through the total number of aminacid residues in the dataset. The combinatorial factor $4!/\prod_a^M t_a!$ corrects for the underestimation of the expected frequency of quadruples with replicated residues due to permutability, M is the number of distinct residue types in a quadruple and t_i is the number of aminoacids of type i. Applying the described procedure to a set of protein structures yields a potential of mean force. Utilizing this obtained potential the energy of a sequence S on a fold ψ can be calculated:

$$W(S,\psi) = \sum_{contacts} q_{contact}$$

Unfortunately it is impossible to normalize the energies obtained in this way because the determination of the ground state energy of S would require to solve the protein folding problem. To avoid this problem and to obtain a comparable property a relative quantity called z-score is defined, which resembles the distance between the energy of S on ψ and an mean energy of misfolds normalized with the standard deviation of the energy distribution of misfolds.

$$z(x,\psi) = \frac{W(x,\psi) - \overline{W}(x)}{\sigma_{W(x)}} \tag{3}$$

The distribution of misfolds is obtained by sliding the test sequenence over an artificial polyprotein, which originated from linking all or at least a main part of the structures in the calibration set together. A polyprotein of length L allows the construction of $L - l$ missfolds if l is the length of the test sequence. This normalization strategy is only sufficient if it is not necessary to have gaps in the sequence to structure alignment, which is one of the greatest limitations of the Extended Tropsha Potential.

3.3 Walks on Landscapes

A landscape is formed by a set of configurations V, a cost or fitness function $f : V \longrightarrow \mathbb{R}$ and an algebraic structure χ which turns the set V into the configuration space (V, χ). For our purposes the configuration set is the set of nucleotide or aminoacid sequences of a defined length n. The protein potential discussed in the section above acts as the fitness function. The algebraic structure is the move set of the walks discussed lateron. In contrast to section 2 the move set here is pure point mutation. In this case, it is possible to represent the configuration space in form of a schlicht graph. Landscapes on graphs are well studied, e.g. in [63].

Tropinverse - an Adaptive Walk Tropinverse is a simple adaptive walk, used to generate inverse sequences compatible with a particular protein fold. The tool uses the extended Tropsha potential as a fitness function the onlz allowed moves are as said already point mutations. Tropinverse starts with a random sequence and consequently mutates random positions. A mutation is only accepted if the zscore of the mutant is better than the zscore of the current sequence. The algorithm stops if whithin a certain number of tries no better sequence is found.

Tropnewt - a Neutral Walk The program was used to generate sequences with a specified distance to a start sequence and the same fitness as the start sequence. The tool succesively mutates single, random positions in the start sequence and checks the zscore of the mutant. If the zscore is within a certain interval around the start sequence's zscore the mutation is accepted. To prevent back mutations mutations are only allowed at so far untouched positions. If a favourable mutation cannot be found within a specified number of steps, the algorithm discards all mutations made so far and starts again with the original sequence.

3.4 The Virtual CSTR

The purification of the replicase of $Q\beta$ - a bacteriophage with a very small genome - in the 1960ies initiated an avalanche of experiments on evolution studies on a molecular level. The replicase was kept multiplying its own gene in a reactor provided with all educts necessary for the replication reaction. Analysing aliquots of the reactor content in regular intervals made it possible to study evolution in nearly molecular resolution, selecting for the fastest replicase. Nearly, because only higher concentrated species were detectable. Such experiments were ususally performed either in a batch reactor, transferring an aliquot of the reactor's content into a "new" reactor to supply enough educts for replication, an approach called serial transfer reactor or in flow reactors.

In a flow reactor a constant influx of buffer enriched with all compounds necessary for replication makes a transfer into a new tank obsolete. Constant outflow at the same rate as influx maintaines constant volume in the reactor. To provide equal conditions in all parts of the tank it is stirred. The selection process of the reactor is very graphic: only replicases doing their job fast enough will stay in the reactor, the slow ones will simply be rinsed out.
Inventing an algorithm resembling such a flow reactor in silico gives a powerful tool for the computersimulation of evolutionary processes. Simulating evolution on the computer has some important advantages. One is able to trace the life of every single molecule ever existing in the reactor and any interaction with other items can be recorded. All parameters can be easily controlled and processes lasting in reality for decades can be simulated in hours or days. The main disadvantage is that the simulation is highly artificial and results might be far from reality.
The reactor is modeled using differential equations describing the following chemical reactions:

$$
\begin{aligned}
* &\xrightarrow{r(t)} A \\
A &\xrightarrow{r(t)} 0 \\
\ell A + I_k &\xrightarrow{f_k([A])Q_{kk}} 2I_k \\
\ell A + I_j &\xrightarrow{f_j([A])Q_{kj}} I_k + I_j \\
I_k &\xrightarrow{r(t)} 0
\end{aligned}
$$

The rate of influx and outflux of monomeres A depends on the current reactivity $r(t) = \sum_{j=1}^{n} f_j x_j(t)$. The current reactivity is time dependend and is the sum over all soecies j in the reactor of the species' fitness f_j weighed with the species' concentration share $x_j(t)$ in the reactor. The probability of an erroneous replication Q_{kj} with distance d^h_{kj} is a function of the replication accuracy per base q or the mutation rate $p = 1 - q$ respectively and the sequence lenght l:

$$Q_{kj} = q^l \left(\frac{1-q}{q}\right)^{d^h_{kj}} = (1-p)^l \left(\frac{p}{1-p}\right)^{d^h_{kj}}$$

Each replication procedure consumes l monomeres A and produces either an exact copy of the template I_k or a mutant I_j resulting in $2I_k$ or $(I_k + I_j$ respectively.

3.4.1 Simulating Intron Development - The Recombination-Reactor

As described in the section above the whole problem of simulating evolution in a flow reactor boils down to the simulations of a system of coupled chemical reactions. This is exactly the purpose of an algorithm developed by Daniel T. Gillespie in the late 1970ies. This algorithm is capable of making timesteps of variable length depending on the cumulated reactivity in the reactor. This feature means that time steps are short if the reactivity in the reactor is high and consequently time steps are long during phases of low reactivity.
The probability of a particular replication reaction to occur is obtained by: $\text{prob}(S, rep) = k_S^{rep} n_S$, where k_S^{rep} is the rate constant of the replication reaction of S and n_S is the number of substrate molecules S in the reactor. k_S^{rep} is equal to the fitness f_S of the sequence S. For the reactors used during this work usually holds:

$$k_S^{rep} = f_S = z(S, \psi)$$

with the only exception of reactors with enforced selection pressure, where the fitness is an exponential of the zscore. So the fitness usually equals the zscore of the sequence S on a given target protein structure ψ.

The replication value $R(t)$ of the reactor at time t is the sum over all individual's rate constants for replication.

$$R(t) = \sum_{j=1}^{N(t)} k_{S_j}^{rep}$$

The outflow depends on the proportion between the actual number of individuals in the reactor $N(t)$ and the default population size N_{def} and the current replication reactivity $R(t)$ weighted by the actual population size $N(t)$.

$$k_S^{out} = \frac{R(t)}{N(t)} \cdot \frac{N(t)}{N_{def}} = \frac{R(t)}{N_{def}}$$

If the actual populationsize is smaller than N_{def} the outflow channels are less likeley, if $N(t)$ exceeds N_{def} more sequences are removed on average. This relation causes the population size in the reactor to fluctuate around N_{def} with a standard deviation of $\sqrt{N_{set}}$. The outflow value is the sum over all individuals outflow constants:

$$O(t) = \sum_{j=1}^{N(t)} k_{S_j}^{out} = N(t) \cdot k_S^{out} = \frac{N(t)}{N_{def}} \cdot R(t)$$

The current reactivity $A(t) = R(t) + O(t)$ defines the Interval $]0..A(t)]$ for the pseudo random number, which selects the replication or outflow channel for the next step.

4 Computer Simulations

4.1 Determining the Properties of Recombination

Is the assumption, made in the introduction about the effect of recombination true? We are going to investigate the impact of recombination on the expression of two different phenotypes, on the formation of RNA secondary structure and on the ability of amino acid sequences to fold into a specified structure. We compare the fitness of offspring populations created by recombination at a specified position, with the fitness of the parent population, in the hope to identify positions of different susceptibility for recombination.

4.1.1 The Impact of Recombination on RNA Secondary Structures

To study the effect of crossover on the conservation of RNA secondary structure, in each case a set of 30 independent RNA sequences with a common minimum free energy structure were generated with the tool `RNAinverse`. For this set of sequences all recombinants $(30 \cdot 29)$ were generated for each sequence position. The minimum free energy structures of all generated recombinants were calculated with the Program `RNAfold`. The impact of a recombination event was observed by counting the number of recombinants, which retained the original secondary structure, for each crossover position, d_{nr}. This measure, however, is not very meaningful for large complex RNA molecules, because, except for terminal crossover positions, no recombinants retain the original structure at all. Therefore we computed the base pair distance d_{bp} between each recombinant's mfe structure and the original mfe structure. The base pair distance is computed, by setting up a triangle matrix $\mathbf{P}$, called pairing matrix, or pairing table, which is defined as:

$$\mathbf{P}_{ij} = \begin{cases} 1 \quad \text{if} \quad i, j & \text{form a base-pair} \\ 0 & \text{otherwise}\,. \end{cases} \tag{4}$$

The base pair distance between two structures, with the pairing matrices $\mathbf{P}$ and $\mathbf{Q}$ is then obtained as: $d_{bp} = \sum_{ij}(|\mathbf{P}_{ij} - \mathbf{Q}_{ij}|)$. To determine the effect of recombination in more complex RNA molecules, we calculate the mean, variance and the minimal d_{bp} of the mfe structures of the set of recombinants, obtained by crossover at a specified position. This measure proved to be a good indicator for positions that are less susceptible to recombination in large RNA molecules. All computer programs mentioned so far for RNA experiments were taken from the *Vienna RNA Package* [32, 74]. To simplify the comparison between the impact of recombination at a specified position and the secondary structure at the position, the secondary structure is represented in all following charts as a mountain plot. The mountain plot function m starts at a value of zero and is increased by 1 unit at positions with an "opening" base pair. At unpaired positions m remains unchanged, at a position with a "closing" base pair, m is decreased by 1 unit.

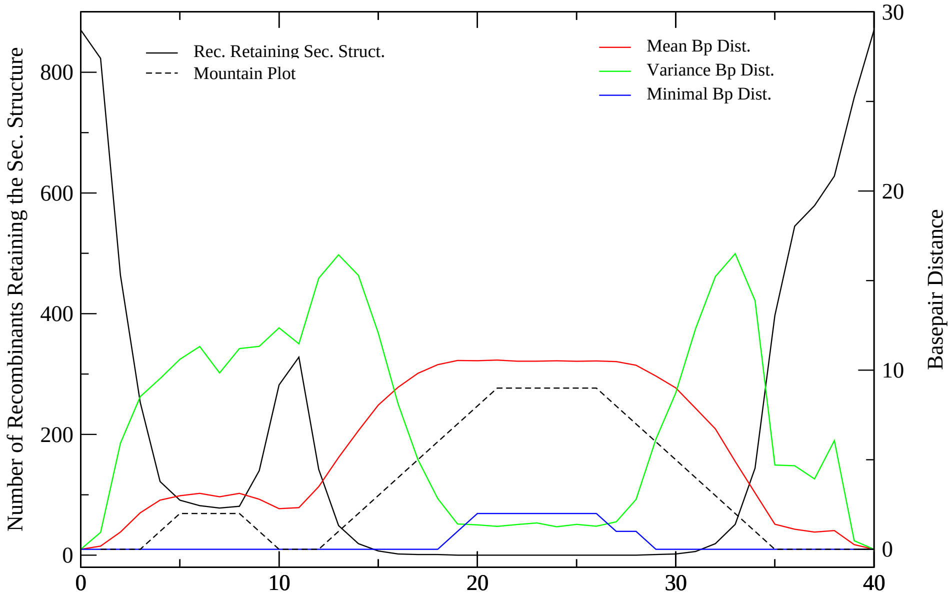


Figure 7: Crossover position dependent RNA sec. structure conservation

The curve of the number of recombinants retaining the original structure, d_{nr}, has in all cases a trough-like shape, this is steeper than we expected, assuming that recombination events at terminal positions affect the formation of structure less than at positions in the core. Obviously, the effect of recombination in a stack region is even predominantly negative, when the stack is placed near the termini. Figure 7 shows the impact of recombination on the structure formation of a small artificial RNA structure. The hinge region is obviously less susceptible to crossover compared to stack or loop regions, almost one third of the recombinants retains the original structure. In the stack and loop regions on the contrary, d_{nr} lies between 0% and 10%

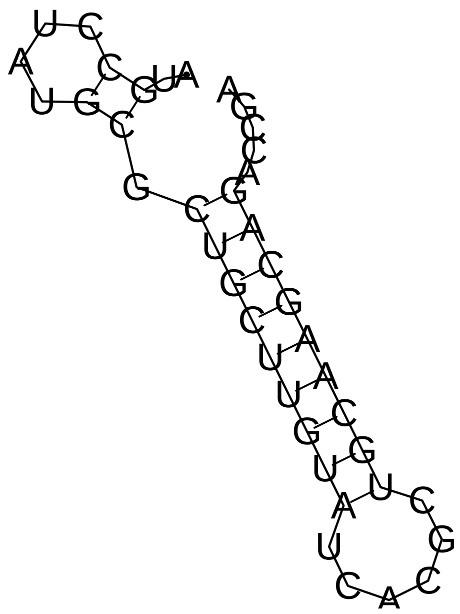


Figure 8: Secondary Structure of the artificial RNA sequence.

of the recombinants. The curve of the minimal d_{bp} is not very meaningful in this example, because at most sequence positions, at least one recombinant retains the original structure, resulting in a minimal d_{bp} of zero. Mean and variance of d_{bp} agree with the behaviour of d_{nr}, but are not of special interest.

In a more complex molecule, like the phenyl tRNA (cf. Figure 9), the number of offspring retaining the structure, d_{nr}, is of no use to distinguish positions of less susceptibility to recombination. The d_{nr} curve falls steeply to values around zero, neutral recombination events are found only at the very terminal positions. The most meaningful curve in this experiment is the minimal d_{bp} curve, which correlates nicely with blocks of secondary structure. Recombination events at all positions between stacks on the central multi loop of the tRNA structure, produce a small number of offspring with the original mfe structure. This fraction is too small, to be visible in the d_{nr} curve, but is nicely reflected in the minimal d_{bp} curve, which is of course zero at such positions. In the case of the tRNA structure, the variance curve of d_{bp} correlates well with the secondary structure, the mean of d_{bp} in contrast drops only slightly at the particular regions, as the main fraction of recombinants does not retain the original structure and is obviously quite distant, as the mean d_{bp} remains at a level at least twice as high as the minimal d_{bp} in stack regions.

Large RNA structures like the mfe structures of 5S rRNA (cf. Figure 11) or an auto-catalytic RNA (cf. Figure 13 on page 44), show a more complex reaction to recombination. In case of 5s rRNA, there is no simple relationship between d_{nr}, minimal or mean d_{bp} and the secondary structure. The variance correlates to some extent with the secondary structure, however this correlation is not strong enough to decide which positions in this structure are especially insensitive to recombination. Surprisingly in the case of the structure of a self replicating RNA species, the d_{nr}, mean and minimal d_{bp} curves reflect the secondary structure. A crossover at the 5' terminal joint position yields a fraction of 50% of recombinants, which retain the original structure. The correlation between the mean d_{bp} plot and the secondary structure is significant, the same applies for the minimal d_{bp}. The self replicating RNA is 3 times as long as the 5S rRNA. This suggest that the loss of correlation between the original secondary structure and the number of recombinants retaining this structure, as well as mean and minimal d_{bp}, cannot be related to the length of the structure, but rather to its complexity.

Species, with the RNA secondary structure as the observed phenotype, have positions which are less or even hardly susceptible to a recombination event at this position. The correlation between the secondary structures and measures of the impact of recombination decreases with increasing length and complexity of the investigated structures. This can be related to the steep increase of possible structures with the length of the molecule, which increases approximately with 2.6^n, if n is the length of the sequence. Probably sequences which fold into a complex mfe structure are in the minority in the sequence space. This could explain, why the recombination experiment with a complex structure of medium length yields less correlation than a much longer simple structure. The chance to create a sequence with a particular complex mfe structure through recombination is small.

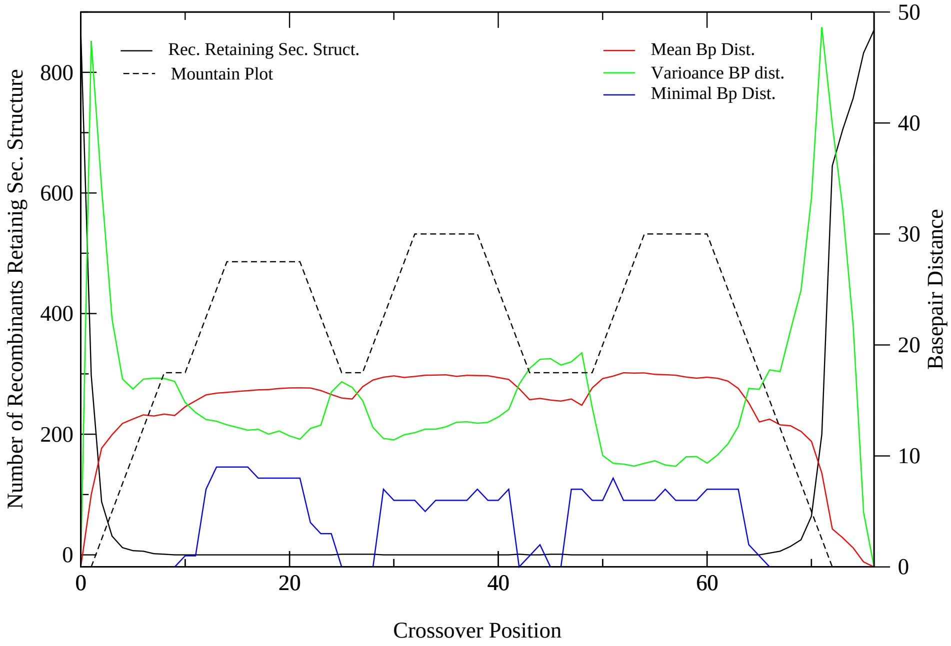


Figure 9: Phenyl tRNA (4TNA)

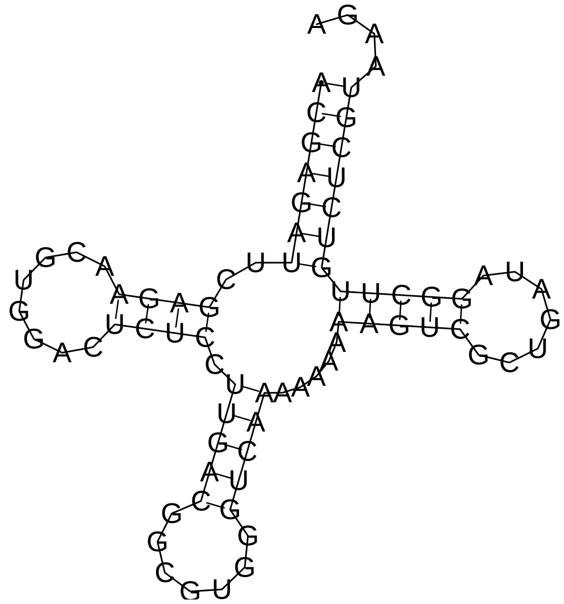


Figure 10: Secondary Structure of the Phenyl tRNA (4TNA)

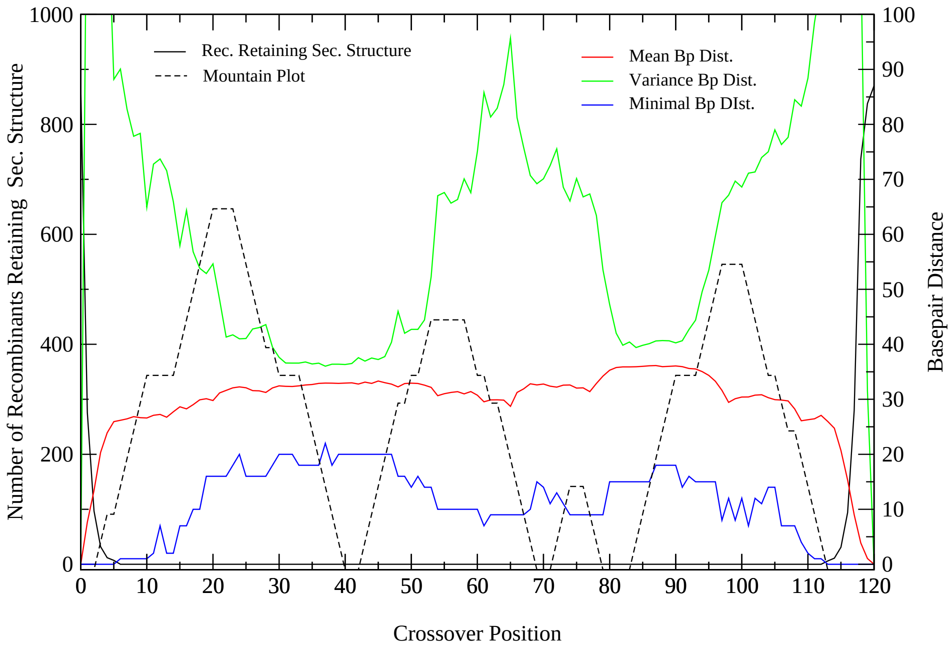


Figure 11: 5s rRNA Ginko Biloba

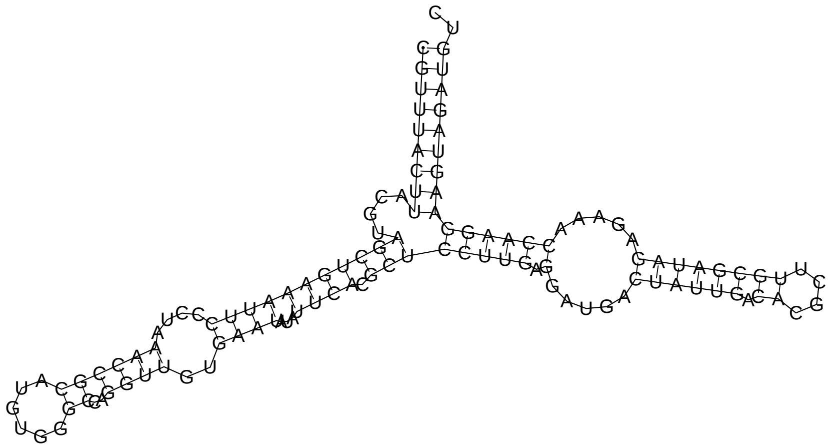


Figure 12: Secondary Structure of 5S rRNA Ginko Biloba

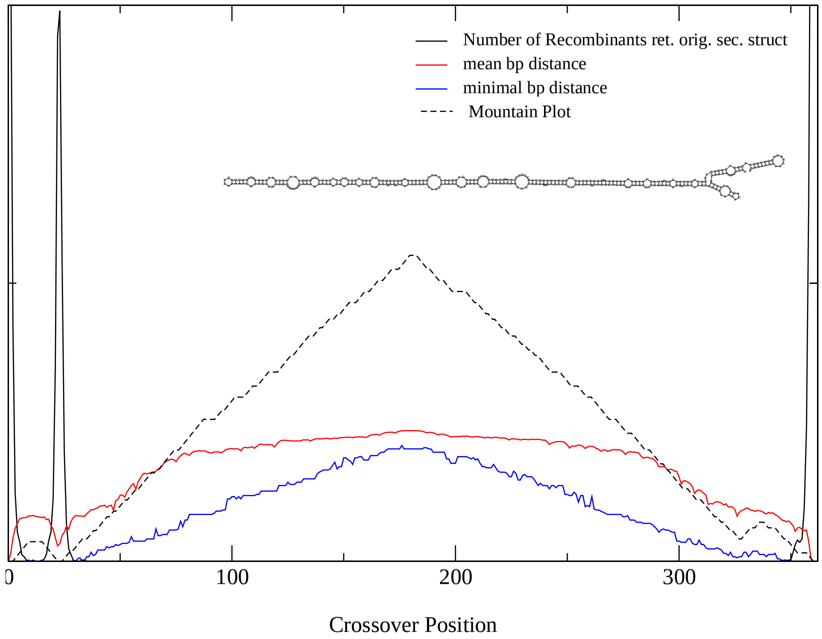


Figure 13: A self-replicating RNA

4.1.2 Recombination of Protein Genes

Analogous to the RNA recombination experiments a set of 30 independent amino acid sequences compatible with a given protein structure was created with the program `Tropinverse`. This tool is based on a knowledge based protein potential [69, 70], described in chapter 3.2.2. It computes amino acid sequences compatible with a given protein structure (pdb file), by performing an adaptive walk starting with a random sequence. A `Tropinverse` run yields a series of sequences with increasing z-score with each pair of consecutive sequences differing in one point mutation.

A startset for the recombination experiments with n sequences was produced by extracting the first sequence having a z-score two units better than the z-score of the native protein's sequence out of n different `Tropinverse` series. In consequence all recombinants between any pair of sequences were computed for each sequence position. The effect of recombination is quantified by

$$\Delta z(x, \psi) = z(x, \psi) - \frac{1}{2} \cdot (z(a, \psi) + z(b, \psi))$$

which measures the ability of x to fold into ψ relative to its parents a and b. $z(x, \psi)$ is the z-score of a sequence x on the fold ψ which is defined by equation 3 on page 33. The distribution of Δz is computed for recombination events at each sequence position. In Figure 14 we show mean and standard deviation as a function of the position of the recombination point.

The mean $\Delta z(x, \psi)$ curve in this figure exhibits the expected v-like trend, we proposed to find in the introduction. An unfavourable recombination in the terminal regions of the protein gene will affect the protein fold much less than a recombination in its core region. Taking into account that most recombination events in the protein gene are rather unfavourable to the expression of the correct structure one should expect a funnel like curve for the average $\Delta z(x, \psi)$. Compared to the analogous experiment with RNA secondary structure formation, we find some apparent differences between the curves. The mean $\Delta z(x, \psi)$ is relatively flat, compared to the d_{nr} curve (number of recombinants retaining the original structure) or the mean d_{bp} (base pair distance) curve. The RNA curves are smooth, whereas the protein curves are rugged. The peaks in the RNA curve sharp and distinct, while the protein curves exhibit very broad peaks. This discrepancy reflects

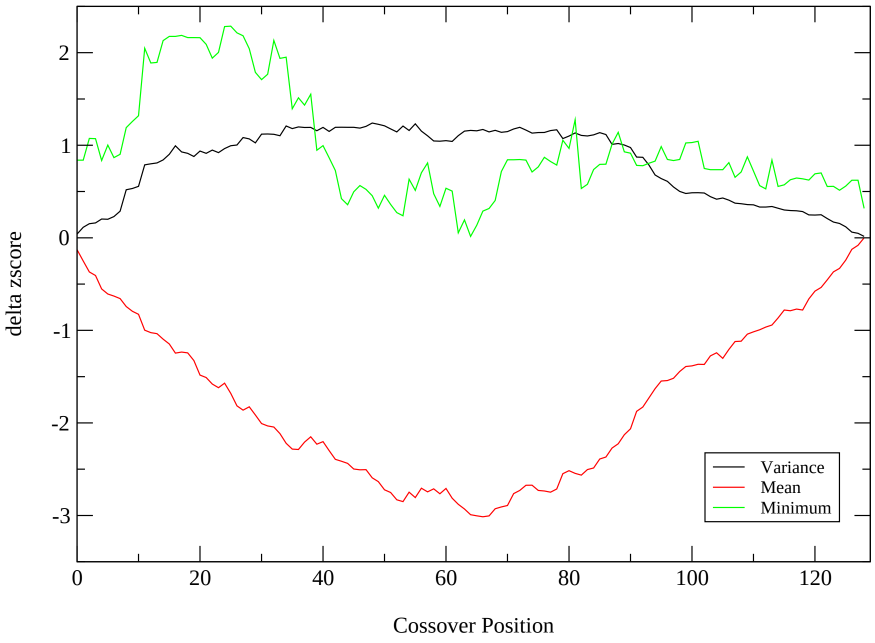


Figure 14: The crossover-position dependent impact of recombination on a protein gene.

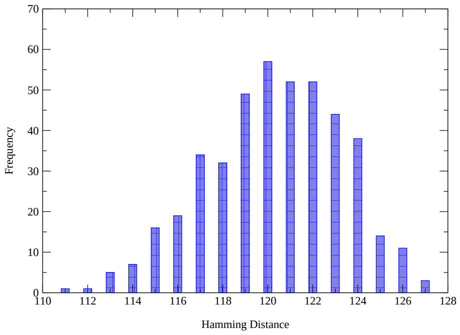
the fundamental differences between the fitness functions used. The interactions in the RNA secondary structure energy model are binary and exclusive, i.e. a base-pair is either formed or not, without any intermediate forms and each nucleotide can participate in one base-pair at most. On the other hand residues in proteins interact with many other residues. The corresponding interaction energies may vary significantly and there is in general not a single dominating contact. The exchange of residues via recombination in an RNA molecule either destroys the base-pairing pattern or leaves it intact. In a protein the interactions with other residues can be affected with a wide range of consequences. The probability that the phenotype remains completely unchanged by recombination is small, on the other hand the probability that the phenotype is destroyed completely is much lower than for RNA.

Though most local peaks in the mean curve correspond to turn or coil regions in the protein structure, they clearly do so in a not significant way. Therefore calculating the position-wise mean delta z-score of all recombinants is not a suitable means to predict module boundaries. The “Minimum” curve in figure 14 shows the z-score of the best recombinant for one crossover position.

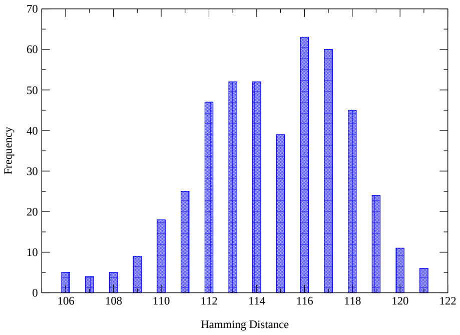
While it looks promising, the curve is not of particular relevance for this experiment but can be some help for understanding the following CSTR simulations. The recombinant with the best z-score has the greatest chance of all sequences in the reactor to produce offspring. A correlation between this curve and the protein structure elements is not obvious.

Distance Dependent Recombination Statistics The experiments described so far about recombination in protein genes were not completely satisfying, as we found only the expected funnel-like shape of the fitness versus crossover position curves, but not the proposed prominent positions of reduced sensitivity to recombination. Probably the experimental setup was too artificial. We presumed that the idea of completely unrelated sequences might have been wrong because most sequences that recombine frequently in nature might be more related due to a common evolutionary origin. Consequently we created an experiment to explore the influence of different grades of homology in the population on the impact of recombination. In the following computation we repeated the previous experiment, but instead of using a set of completely unrelated sequences we created sequence sets with a defined hamming distance between a start sequence and the rest. The sequence sets were generated with the program `Tropnewt` which performs a neutral walk. The same start sequence – one out of the compatible sequence set of the previous experiment – was used for all `Tropnewt` runs. `Tropnewt` generates a valid sequence S with distance d to the start sequence by performing d accepted moves with point mutations allowed as the only move. After each move $z(S, \psi)$ is calculated and the move only accepted if $z(S, \psi)$ lies within a tolerance region around the parent sequences z-score. However for the effect of recombination the homologies between the sequences are important rather than the distance to a start sequence. Therefore we calculated the pairwise distance distribution in a set. The histograms of the distributions are shown in figure 15 on page 48.

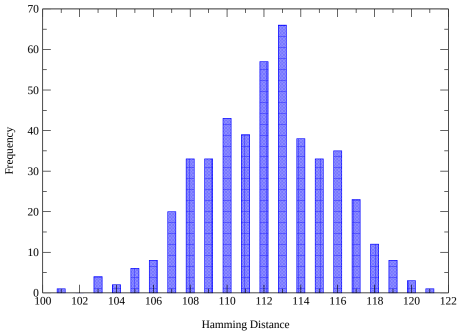
Figure 16 gives a comparison of all mean Δz curves, ranging from sets with distance 20 to sets with distance 120. The curve entitled "dist129" is the result of the previous experiment acting as a reference in this chart. Certainly recombination between sequences very similar to each other will not have a great effect this is reflected by the curves "dist20" and "dist40" with approximately 84% homology and 68% respectively. With decreasing homology the effect of recombination becomes more and more noticeable and peaks in the



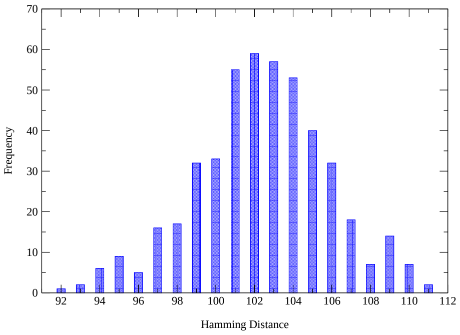
(a) dist129



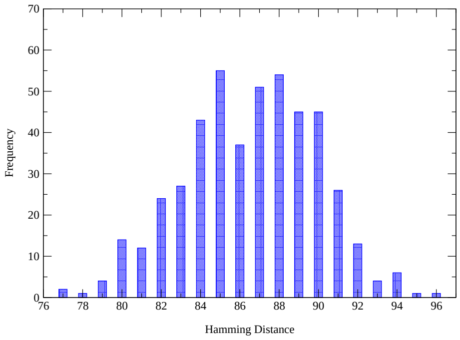
(b) dist120



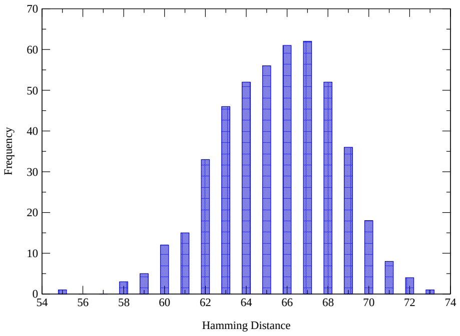
(c) dist100



(d) dist80



(e) dist60



(f) dist40

Figure 15: Pairwise distance distribution of the various startsets for the distance dependent recombination statistics. All sets were generated with the tool `Tropnewt`, using the same start sequence. The number in the name of the startset accounts for the distance of each sequence to the start sequence

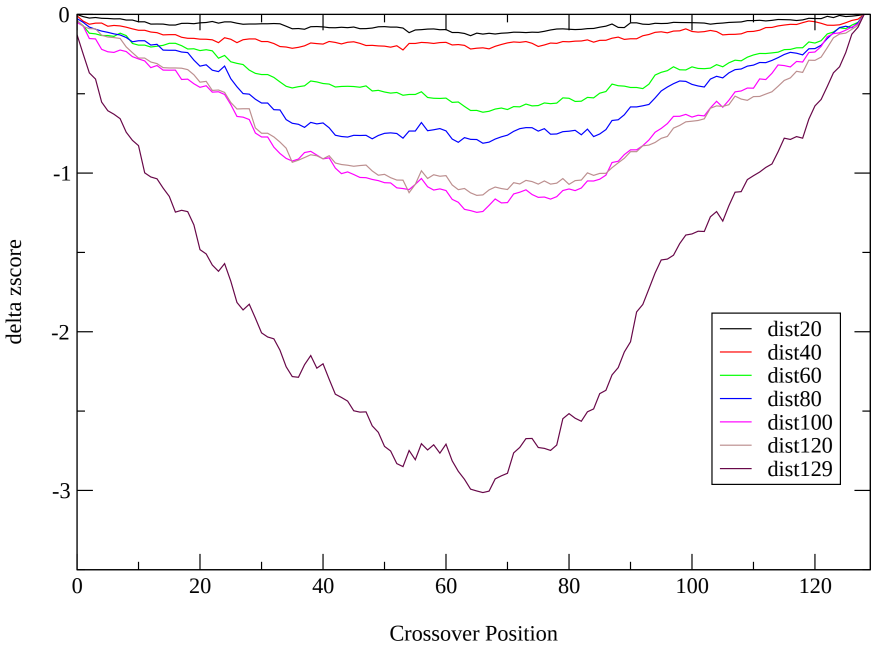


Figure 16: The effect of different grades of homology in the initial population on the impact of recombination: A comparison of mean Δz curves. All populations used originated from the same start sequence by random mutation. Structure:1LSE Startset: mediumzscore (parent z-score 10.015, avg. zs of set 12.9)

curve are forming more and more clearly. All peaks in the various distance curves correspond to peaks in the "dist129" curve. Astonishing, on first sight, is the big gap between the mean Δz of "dist120" and the "dist129" reference curve in figure 16. Why should a small difference in homology affect the z-score of the recombinants that much? The gap can be explained by the way `Tropadapt` computes the sequence set. Making a move in which one amino acid is replaced by a very similar amino acid is much more likely to be accepted than performing a move which e.g.. replaces a hydrophobic amino acid with an acidic one. Consequently all sequences in the "dist120" are much more related because of their common ancestor and a series of rather "cheap" mutations than the sequences in the "dist129" which are completely unrelated because each of them is the product of a genuine adaptive walk starting with a random sequence. A similar chart was computed for the best z-score per sequence position using the same sets as above, see Figure 17 on

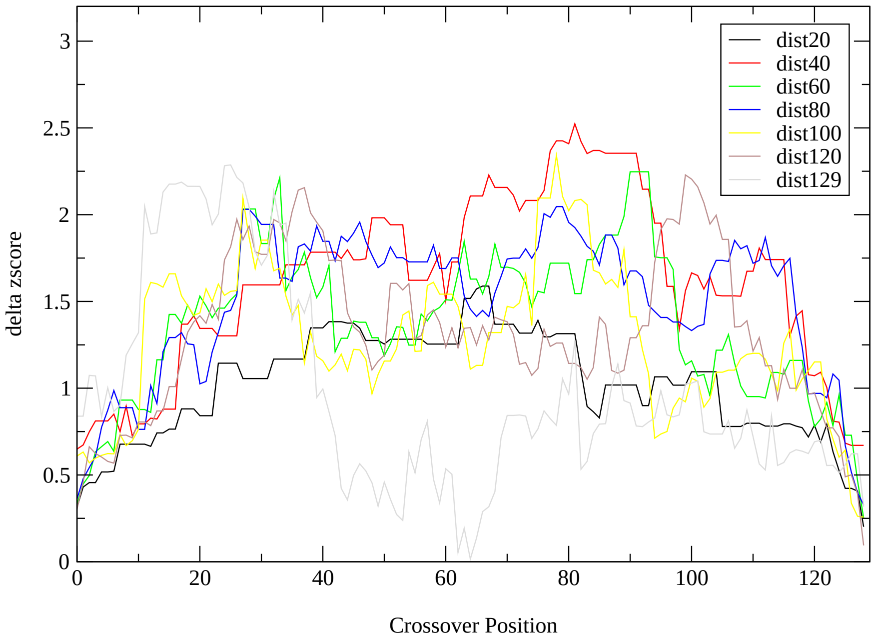


Figure 17: Comparison of minimum Δz curves using sets with different grade of homology.

page 50. The shapes of the curves are complex and do not allow to find any correlation between the decreasing homologies of the start sets and the peaks in the corresponding curves.

We have argued that the requirements for the modularization of genes under recombination are (a) a general negative effect of recombination and (b) the existence of a subset of positions at which recombination is less detrimental. These positions would then form the "module boundaries". We could verify (a), but assumption (b), that such positions exist, preferentially ones that separate building blocks of the protein, which we concluded are the secondary structure elements, did not proof to be true, at least for the fold of 1LSE. Even a modification of the relatedness of the sequences, resulting in all gradations between non-homologous and homologous recombination, does not change this result. In contrast to this, our results with some RNA secondary structures suggest, that our idea of outstanding recombination positions is not entirely false. We speculate that the energy model of the RNA secondary structure prediction, allows or promotes a certain extent of modularization.

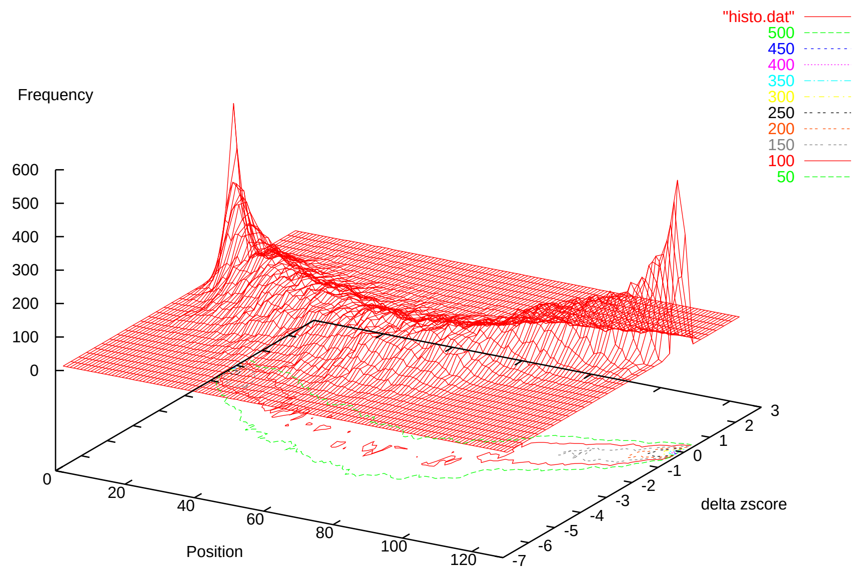


Figure 18: Histogram of the position-wise z-score distribution after recombination of the startset dist129

4.1.3 A Pure Recombination Genetic Algorithm

Though the invention of start sets with a certain homology did not improve the results of the last two experiments, we were interested how this homology affected the next generation. Moreover we wanted to know how reproduction by recombination, without aditional operators, changes the variability in the offspring. The formal analysis of recombination operators revealed, that recombination, if not paired with point mutation, reduces the variability in a population. Additionally one can ask how effective recombination is as an operator in an evolutionary optimization process. To test this a genetic algorithm without any mutations was constructed called `Roundabout`. The routine works very similar to the protein recombination statistics program discussed in section 4.1.2. Using a startset of 30 amino acid sequences generated in the same way as described in 4.1.2 all recombinants for all possible crossover positions were generated and mean, variance and minimum of the delta z-score distribution calculated. Of these recombinants the 30 best were chosen and used as a startset for another run of Roundabout. This procedure was repeated until the set of resulting recombinants was completely homologous.

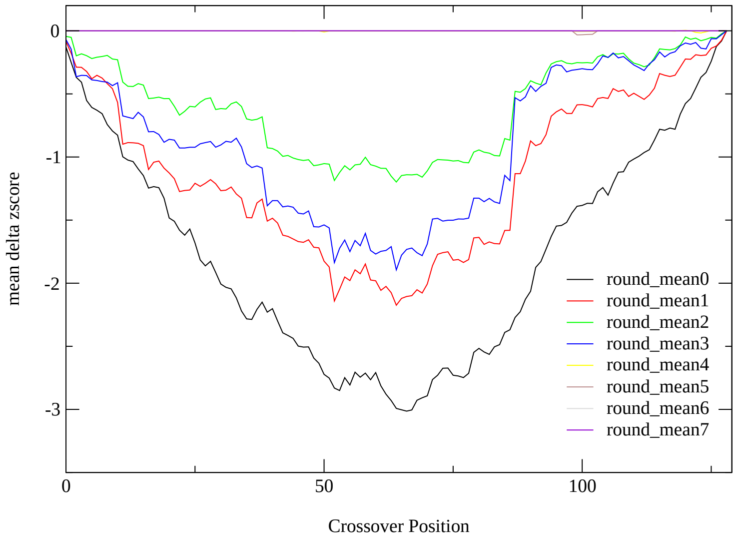


Figure 19: Development of the mean Δz during 7 runs of roundabout

The mean Δz curve of the first run, shown in figure 19, is of course totally identical with the curve of the recombination statistics experiment with maximal sequence heterogeneity, discussed in the previous chapter. The curves of the following generations of offspring differ in a similar way, like the curves of the populations of increasing homology differed in the previous experiment. This behaviour hints that there is indeed a continuous reduction in heterogeneity occurring during this experiment. This is confirmed by the analysis of the variance Δz curves (cf. figure 21). The variance among the offspring z-scores reflects the sequence heterogeneity. Recombination of identical sequences results in zero Δz variance, whereas recombination among completely heterogenous sequences yields maximum z-score variance. Of special interest is figure 20 on page 53, representing the curves of the best Δz in the offspring of a crossover of a specified position. The horizontal parts of the curves indicate that in this regions the best sequence is independent from the recombination position, obviously the sequences are already homologous in this parts. The best Δz curves of the first rounds of the genetic algorithm show that recombination is beneficial, however only occasionally.

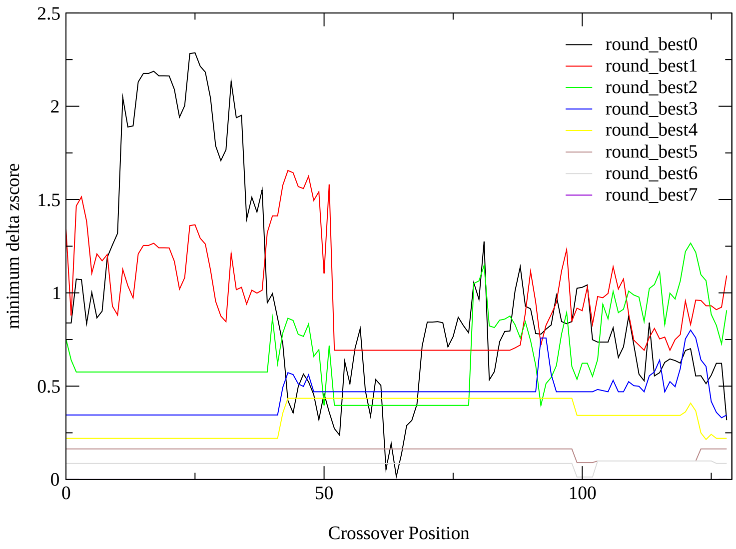


Figure 20: Development of the best Δz during 7 runs of roundabout

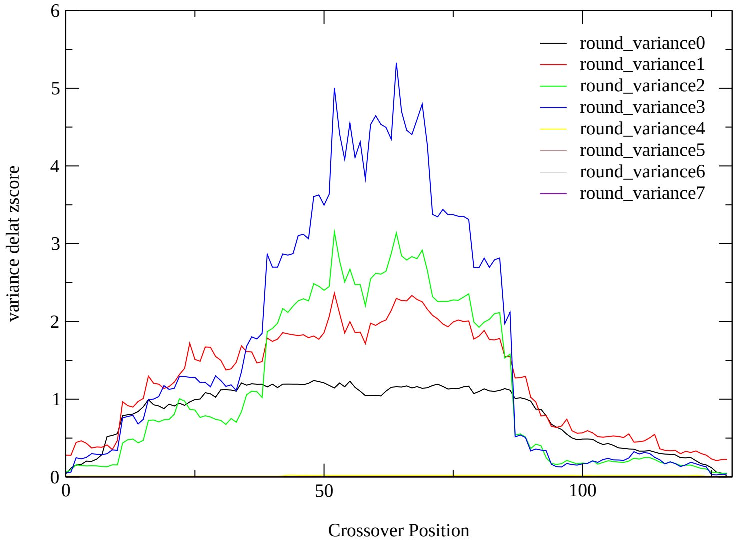


Figure 21: Development of the Δz variance during 7 runs of roundabout

The production of offspring, fitter than its parents, is reduced continuously, with the reduction in sequence variability.

We can summarize, that recombination acts as a homogenising operator in this case. If it works alone, independent of a second genetic editing operator, which introduces a certain amount of variability, like point mutation, the process results in a completely homogeneous population. Recombination is partially advantageous, but depends on some variability to produce fitter offspring compared with the parents. The latter finding signifies that recombination can only be an effective operator in optimization processes, if it is combined with a second operator which keeps the variability at a sufficiently high level. In nature this is achieved by the combination of recombination and mutation.

4.2 Intron development in a Flow Reactor Simulation

If recombination acts as a disturbing force at most sequence positions but is at least neutral or even beneficial at some special positions, an evolutionary simulation of intron placement should reveal these positions. The probability to separate two points on a genome by recombination depends on the distance between both points. Having this in mind one would expect the formation of additional spacers – introns – at positions where recombination is a neutral or even a beneficial event. One would assume positions in the genome to be favoured for intron development which correspond to module boundaries or spots of not clearly defined secondary structure in the protein.

The recombination reactor mimics a continuously stirred tank reactor with reactivity dependent flow rate. The reactor is capable of recombination, point mutation, insertion of special intron characters and deletion of intron characters. The replication probability of each individual depends on its fitness, the probability to die out is constant for all inhabitants, simulating the constant out-flux of the reactor. For a detailed description of the reactor used refer to chapter 3.4 on page 34.

The Intron distribution of 1LSE All recombination reactor experiments were initially done with 1LSE, the crystal structure of a hen egg white lysozyme. Lysozyme is an enzyme with proteolytic activity which can be found in many species in eucaryotic as well as procaryotic kingdom. This made it an interesting candidate for eventual comparison of lysozyme genes of intron bearing and intron-less organisms. For the recombination reactor a set of 900 sequences (30 times the set used for the recombination statistics

Table 1: Parameters of the recombination reactor, used to simulate intron development. At the beginning of each replication the reactor has to decide whether either a recombination or mutations and insertions take place. In case of recombination, a position and a second sequence are chosen at random. In the oppposite case a mutation may occur at each sequence position, the mutation may be an insertion of an intron character, a ppoint mutation, if the nucleotide is not an intron character or a deletion if the nucleotide is an intron character.

mutation rate	0.001	mutations per nucleotide
recombination rate	0.005	per replication
insertion/deletion rate	0.1	per mutation
average population size	1000	

described in chapter 4.1.2) was used as a start set. For the exact parameters of the simulation refer to table 1.

The course of the optimization procedure during a recombination reactor run is shown in Figure 22. In comparison Figure 23 shows the fitness gain during a reactor run without any recombination events at all. Obviously the optimization in this case is not improved by a combination of recombination and point mutation, as the time versus fitness curve without recombination is much steeper, accounting for a faster optimization process. Crossover free optimization even reaches a higher average fitness in total, but considering the fluctuations this is not a significant difference.

The spatial distribution of introns can be depicted by simply counting the intron characters between the coding sequence positions. We remark that n introns at position j means that n introns were counted between j and $j+1$ where j is an element of $[0, 3a]$, with a being the number of amino acid residues of the protein (in case of 1LSE $3 \cdot 129$).

Peaks in the spatial intron distribution plot in Figure 24 correspond mainly with borders of secondary structure elements in the protein structure. In general an intron if invented once is not lost any more, it usually is prolonged during the simulation, only in rare cases introns seem to shrink again. The tendency of elongation of the introns is due to insufficient selection pressure against it at this stage. The probabilities per base for insertion and dele-

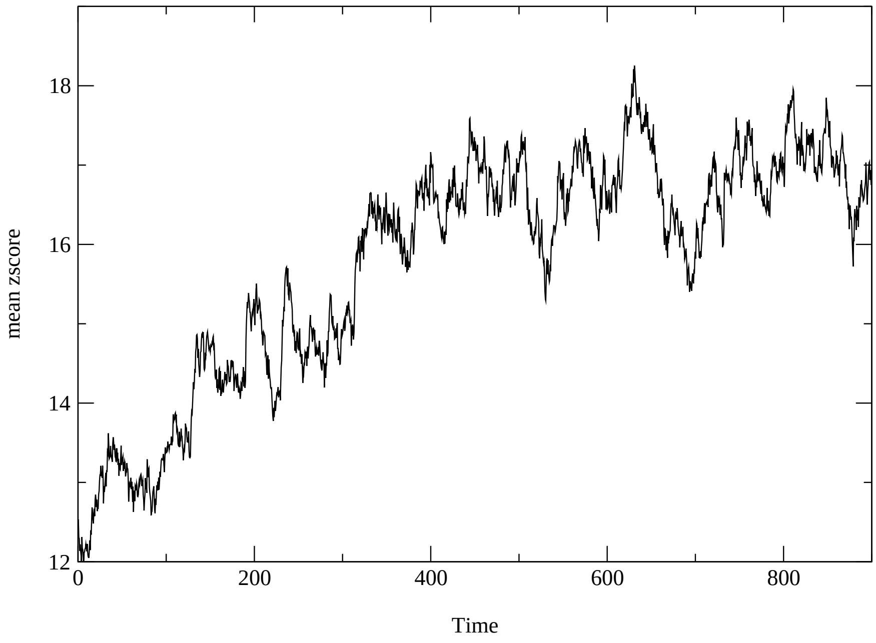


Figure 22: Development of the average fitness during a recombination reactor simulation (Startset Mediumzscore3)

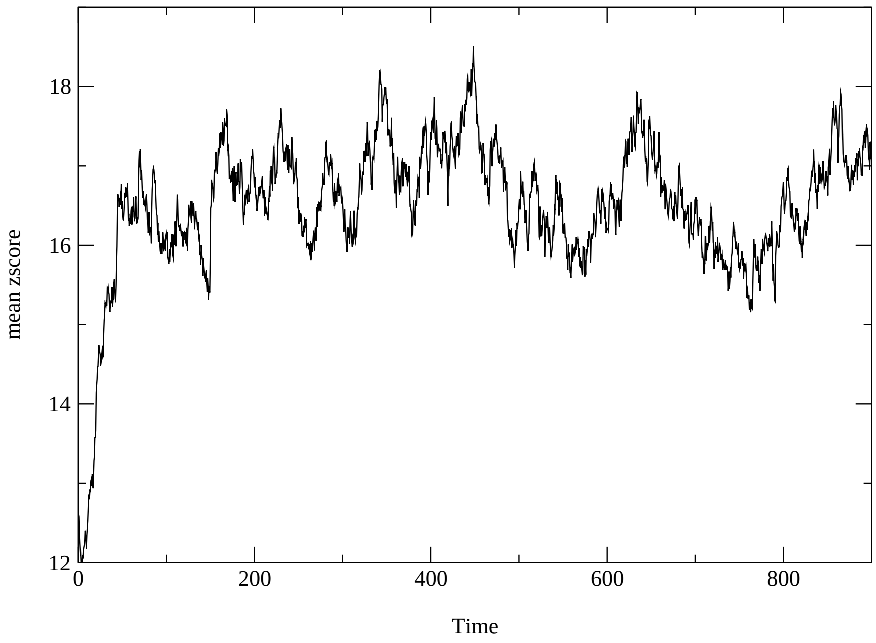


Figure 23: Development of the average fitness during a recombination reactor run without recombination (Startset Mediumzscore3)

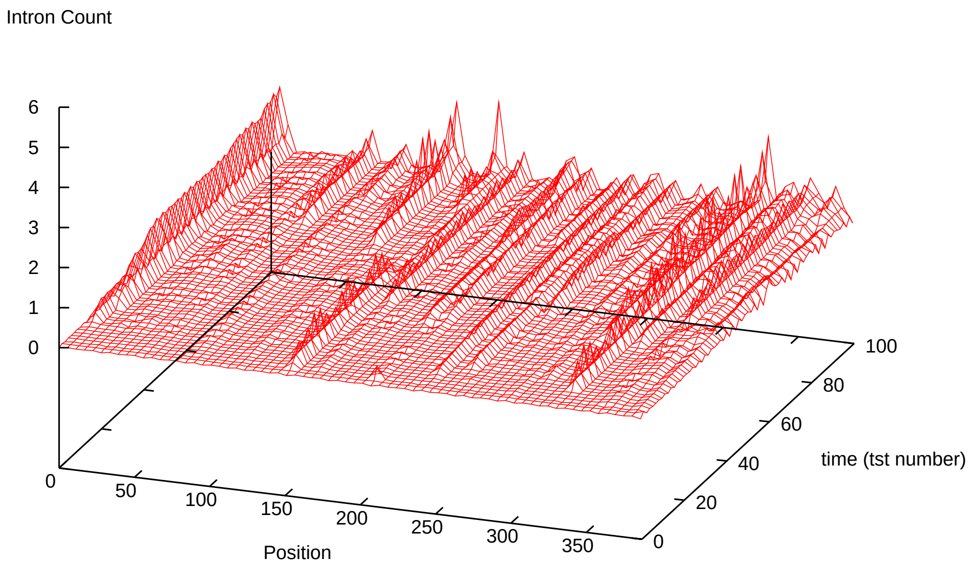


Figure 24: Spatial intron distribution in the 1LSE gene during a recombination reactor run over 100 tst time steps. (Startset Mediumzscore2)

tion during replication are equal. However insertions can happen anywhere whereas deletions can only occur when the affected position is in an intron. Thus the rates of insertion and deletion will become equal only after as many intron characters have been inserted as there are coding characters. This time point lies far beyond the 100th time step in the experiments reported here.

To test whether introns lie preferentially in regions of undefined secondary structure or borders of secondary structure elements, the intron frequency $f^I_{i,j}$ per secondary structure class pair was calculated. Secondary structure class pair means the secondary structure types of the amino acid residue or residues respectively which the intron's confining bases code for. To account for the different incidences of such pairs, the intron count per type pair $n^I_{i,j}$ was normalized through the frequency of occurrence $N^{occ}_{i,j}$ of the pair, which results in $f^I_{i,j} = n^I_{i,j}/N^{occ}_{i,j}$. The secondary structure type per amino acid was extracted from the pdb file using the tool Stride, which is a part of the VMD molecule viewer package.

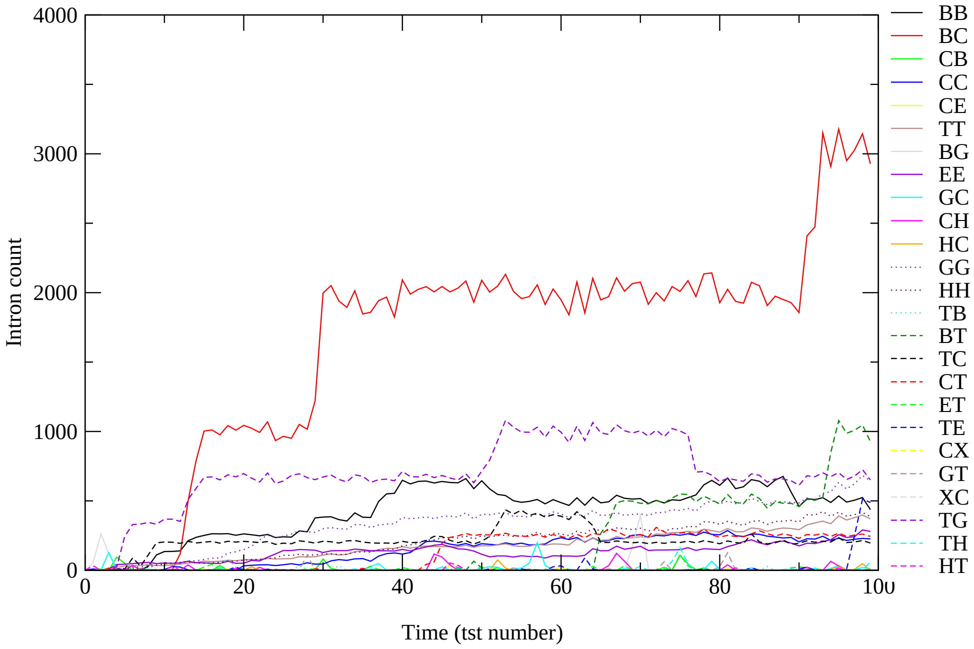


Figure 25: Intron frequency per secondary structure type pair, normalized with the occurrence of each pair (Startset Mediumzscore2)

The run of the normalized intron frequency curves in figure 25 suggests that intron development happens stepwise. If a favourable intron has been invented it spreads fast among the population, which is reflected by the steep steps. For the encoding of secondary structure classes please refer to table 2 on page 59.

A less confusing view of the distribution of introns over secondary structure types is to pool them into three classes of secondary structure types. Type S with a defined secondary structure, type B for secondary structure borders and type N with undefined secondary structure. The intron probability $prob^I(T)$ per class is calculated as follows:

Table 2: Code table for Stride secondary structure class codes

Key	Secondary Structure Class
B	isolated bridge
C	coil
E	strand
G	310 helix
H	alpha helix
T	turn
X	5/ or 3/ end (not a stride character)

$$freq^I(T) = \frac{\sum_{ij\in T} n^I_{ij}}{\sum_{ij\in T} N^{occ}_{ij}}$$

$$prob^I(T) = \frac{freq^I(T)}{\sum_T freq^I(T)}$$

where $T \in \{S, B, N\}$ and $S = \{(H, H), (G, G), (E, E)\}$,
$N = \{\{T, T\}, \{C, C\}, \{T, C\}\}$,
$B = \{\{H, T\}, \{T, G\}, \{T, E\}, \{B, T\}, \{H, C\}, \{G, C\}, \{C, E\}\{B, C\}\}$.

The measure $\mathrm{prob}^I(T)$ is the probability to find an intron in class T. If our assumption that genes modularize under recombination is wrong, the introns should be found with equal probability in all classes, which means that $\mathrm{prob}^I(T) = 1/3$ for all $T \in \{S, B, N\}$. We expect to find a low intron probability in class S, anf higher probabilies for the classes N and B. The development in the first simulation of the intron distribution over secondary structure class types during 100 tst time steps is shown in Figure 26.

This curves look quite promising. After an equilibration time of roughly 50 time steps the probability to find an intron in exactly one of the three

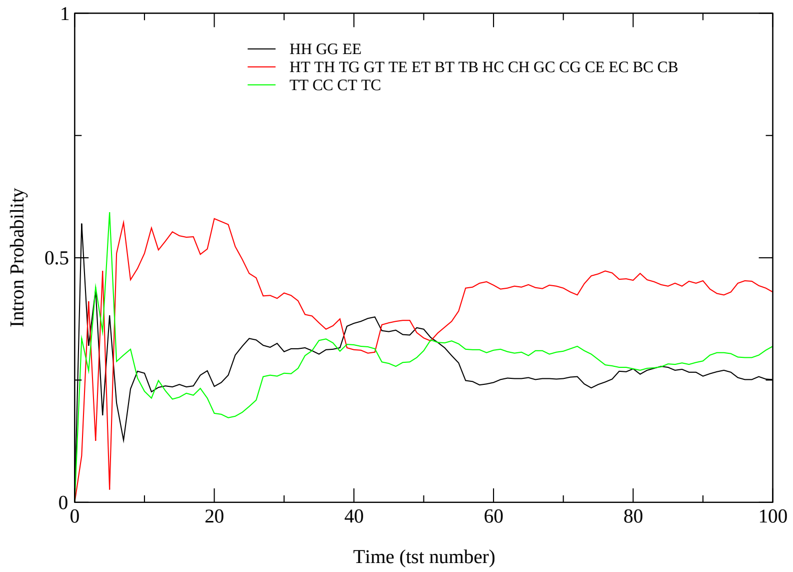


Figure 26: Intron Probability per secondary structure class(Startset Mediumzscore2)

secondary structure classes has become stable. The probability to find an intron at a secondary structure border is nearly two times as high as finding it within a region of defined secondary structure. A recombination event within a turn or coil region should be more beneficial on average than a recombination event within helices or beta sheets but the combination of adjacent amino acids is not completely free in such regions. This is reflected in the closeness of the N curve to the S curve.

Because of this encouraging result, a series of similar reactor experiments were performed with varying startsets (with equal mean z-score) with the structure of 1LSE and subsequently with three different protein folds.

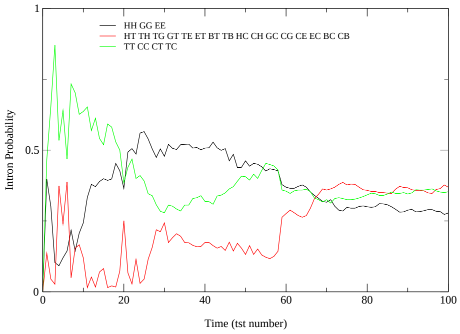
Despite the encouraging start, the obtained intron distributions in several runs with 1LSE as the target structure are not consistent. It seems that the result of the reactor run depends strongly on the start set and on the initialization of the random number generator. Figure 27 on page 62 shows the outcome of these calculations. Considering these results one should be interested in the intron distribution of reactor runs without any recombination, thereby removing the pressure which could lead to protein modularization. This was done by simulating the intron development in the reactor with the usual parameters (cf. table 1 on page 55). The only difference is the recombination probability which is set to zero. The outcome of two such negative control runs is shown in figure 28 on page 63. Recombination appears to cause a larger variation in intron distribution but does not seem to concentrate introns on secondary structure boundaries.

The same procedure as described for the lysozyme fold was applied to the crystal structures of birch pollen allergen betv1, called 1BV1, to the crystal structure of interleukin-2, 1IRL and to a serine esterase known as the charcot leyden protein 1LCL. Figure 29 on page 64 shows the intron distribution curves obtained from three recombination reactor runs with the fold of 1LCL. The picture is more or less the same as in the 1LSE runs. The intron distribution differs greatly between individual runs, on average the introns are not formed preferentially in one of the three secondary structure classes. Figure 30 on page 65 shows the outcome of identical calculations done with the fold of interleukin-2, 1IRL. The intron distribution differs even more between individual runs than in the case of 1LCL. Finally the results of the recombination reactor simulation of 1BV1 are shown in figure 31 on page 66.

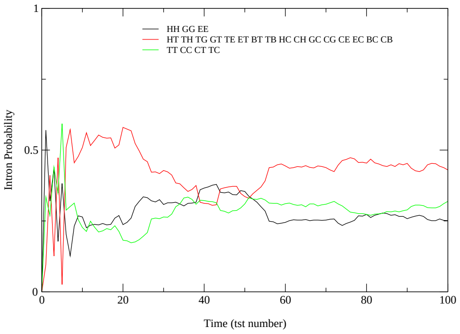
The intron placement in the reactor simulation does not significantly depend on the secondary structure of the corresponding amino acid residues of the protein into which or between which the intron was placed.

A noticeable effect of gene modularization under recombination could be prevented by several reasons: The fitness differences between offspring created by a beneficial recombination event and a detrimental recombination event are relatively small. Thus the fitness function is maybe not sensitive enough to discriminate between beneficial and disadvantageous recombination events, i.e. the fitness function does not provide enough selection pressure for beneficial recombinations and thereby for sequences with introns placed in between secondary structures. This interpretation is confirmed by the fitness versus time plots, shown in figure 22 on page 56. Those curves indicate that the selection pressure is rather low, such that the fitness level cannot be maintained, but the fitness fluctuates over a broad range.

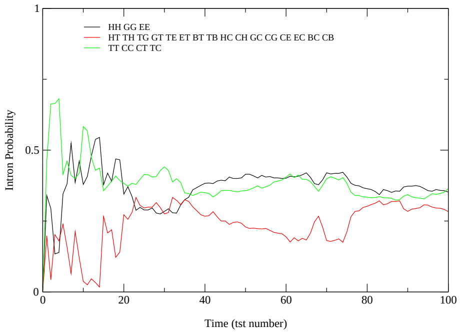
A second feasible explanation is that the majority of sequences are products of mutation rather than recombination events. We have discussed earlier (cf. 4.1.3) that recombination homogenizes the population and that recombination shows only – positive as well as negative – effects if the variability in the population is high enough. Therefore we are forced to keep the mutation rate



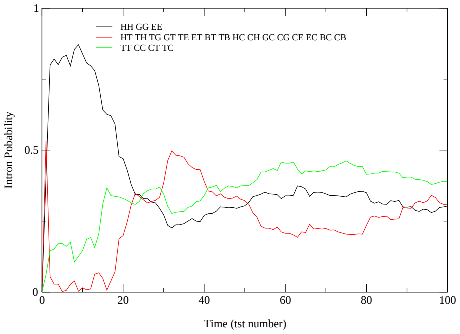
(a) Mediumzscore 1



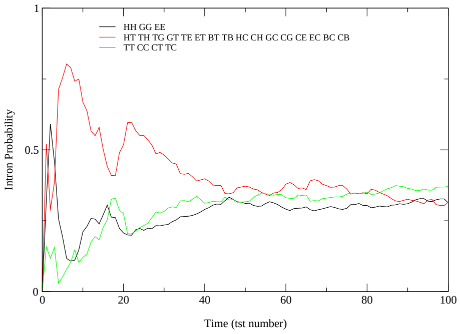
(b) Mediumzscore 2



(c) Mediumzscore 3

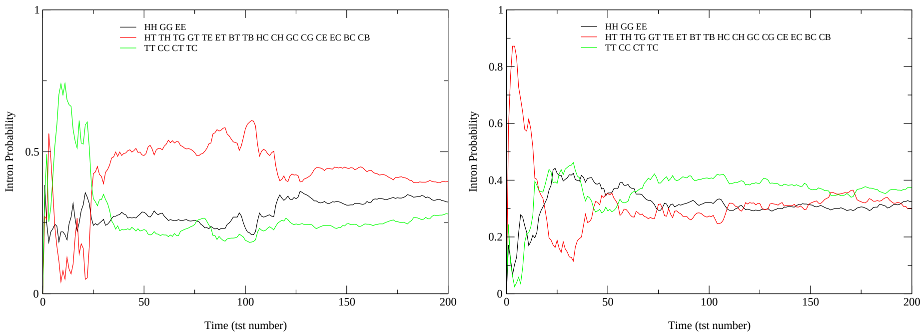


(d) Mediumzscore 4



(e) Mediumzscore 5

Figure 27: Intron distribution in the Lysozyme gene (target structure 1LSE). The Intron distribution was obtained after runs with different startsets. All startsets consisted of 30 different unrelated sequences, each sequence present in 30 copies.



(a) Mediumzscore 2 (b) Mediumzscore 3

Figure 28: Intron distribution without recombination events.

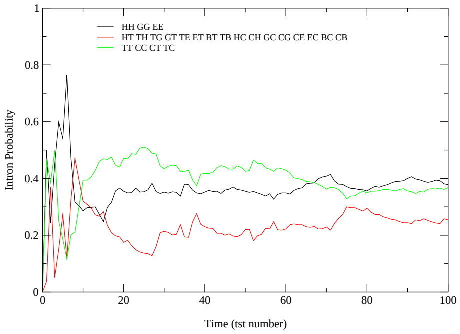
at a relatively high level, compared to the recombination rate, to maintain a sufficient variability. A side-effect of this necessity is that most offspring is created by point mutation and not by recombination. As point mutation does not interfere with introns of the gene, the sequences might not be under a sufficiently strong selection pressure for biased intron placement.

A third explanation for our results is that the fitness function itself does not promote the modularization of genes. In other words the fitness function, which is designed to resemble a specific reaistic selection as much as possible does not have this feature, an therefore does not lead to modularization.

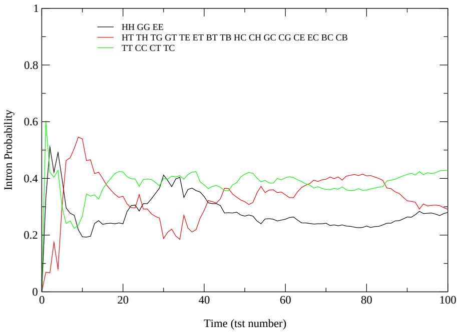
A workaround for the explanation concerning the low recombination frequency due to maintenance of variation would be to increase the population size in the reactor. This would increase the pool of different species and therefore provide more "genetic material" for recombination. Unfortunately we are limited in this point by computer hardware.

A possible workaround for the problem of a weak selection could be to steepen the fitness function, thereby increasing the fitness difference between two differently optimized sequences. We used a reactor which was identical with the ones before, with the only difference that the fitness of a sequence was not its z-score with the target protein fold, instead the fitness $f(S)$ was defined as:

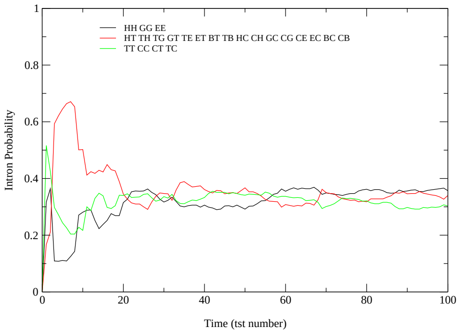
$$f(S) = \exp(z(S, \psi) - z(S_{wt}, \psi))$$



(a) Mediumzscore 1



(b) Mediumzscore 2

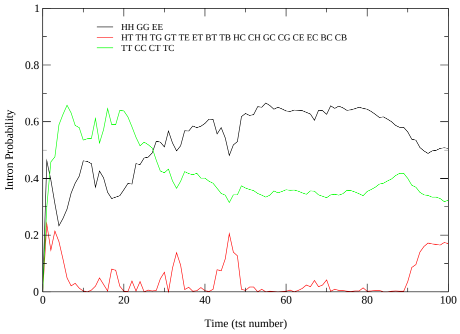


(c) Mediumzscore 3

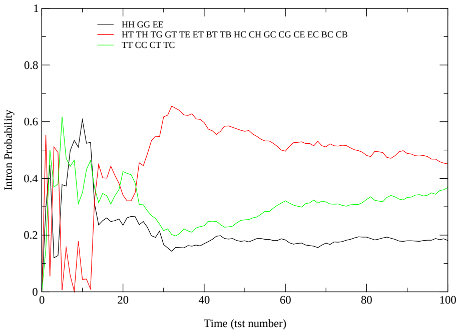
Figure 29: Intron distribution in the 1LCL gene after 100 time steps.

The resulting intron distribution in the lysozyme gene during four different reactor runs with exponential fitness function are shown in figure 32 on page 67.

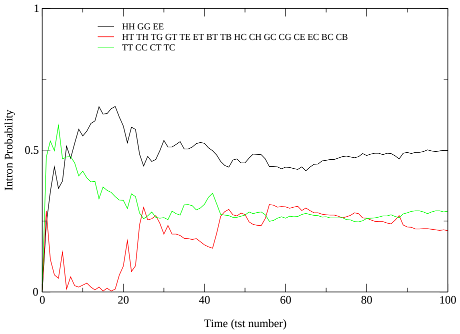
In principle the four curves do not reveal anything new. Compared to the intron distributions of reactor runs with the flat fitness function (but the same start sets), the four different runs seem more consistent. The intron probabilities of the individual classes of secondary structure type pairs scatter all around one third, which is the expected value for equipartition of the introns. As a control, one might to have a look at the fitness development of those runs, whether the different fitness function results in a better conservation of a high fitness level. Figure 33 on page 68 depicts the fitness



(a) Mediumzscore 1



(b) Mediumzscore 2

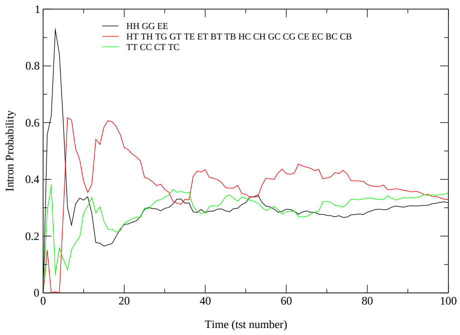


(c) Mediumzscore 3

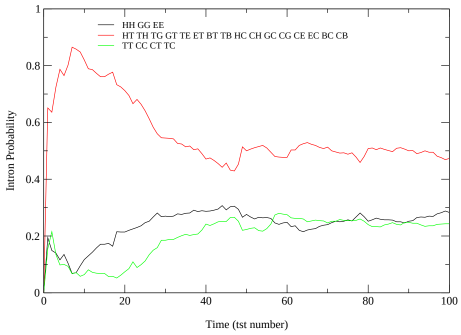
Figure 30: Intron distribution in the 1IRL gene after 100 time steps.

development in the reactor versus the number of replications. The stepwise fitness propagation in the saturation phase is typical for reactor simulations with strong selection.

Additionally we simulated the intron development in a reactor with exponential fitness function, but with out recombination. The fitness during the reactor simulation is compared to the fitness development in the reactor simulation with recombination in figure 33 on page 68. The according intron distribution plots are shown in figure 34 on page 69.

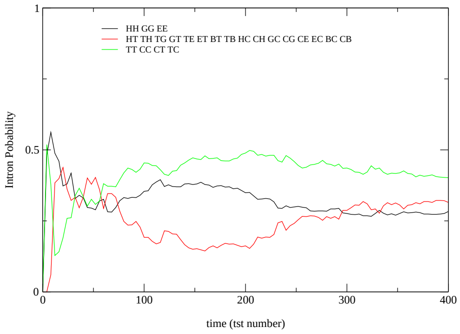


(a) Mediumzscore 1

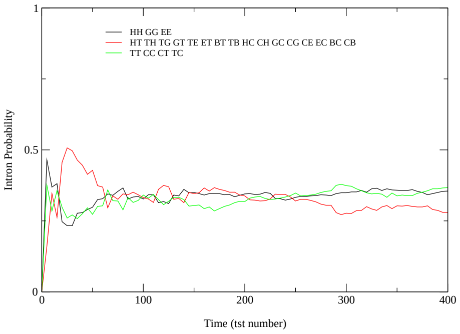


(b) Mediumzscore 2

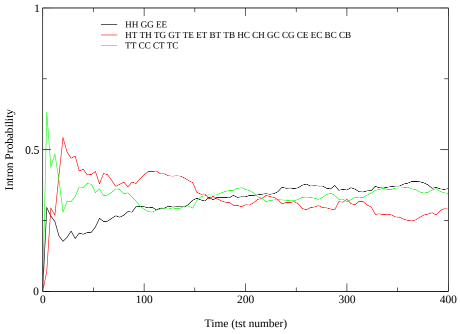
Figure 31: Modularization of the 1BV1 (birch pollen allergen gene).



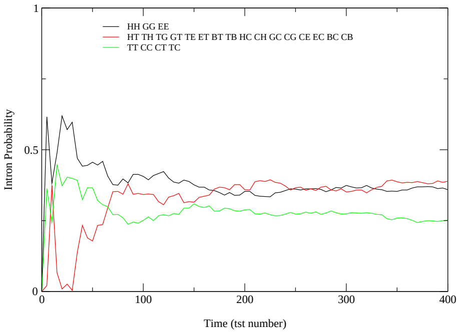
(a) Mediumzscore 2



(b) Mediumzscore 3

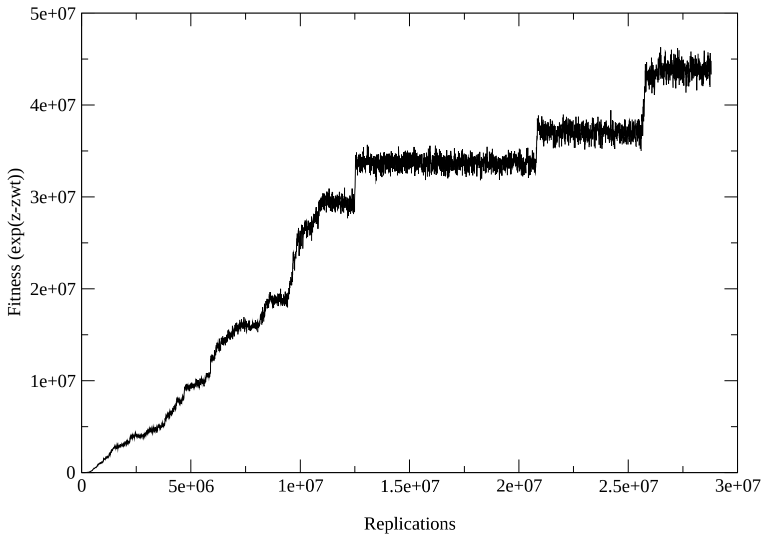


(c) Mediumzscore 4

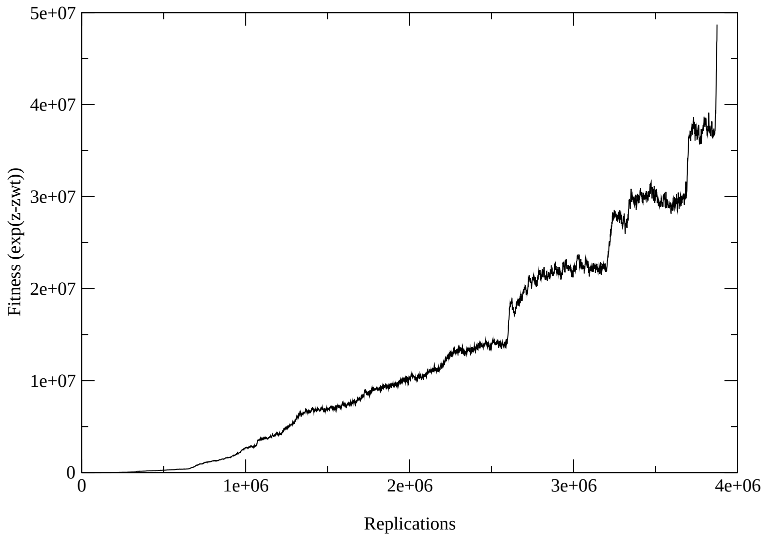


(d) Mediumzscore 5

Figure 32: Distribution of Introns in the lysozyme gene (1LSE) during a recombination reactor simulation, with exponential fitness function ($\exp(z(S,\psi) - z(S_{wt},\psi))$, S_{wt} is the wild type sequence).

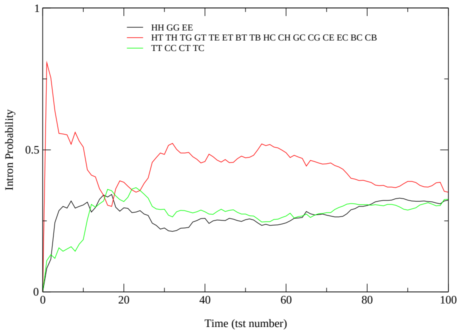


(a) with recombination

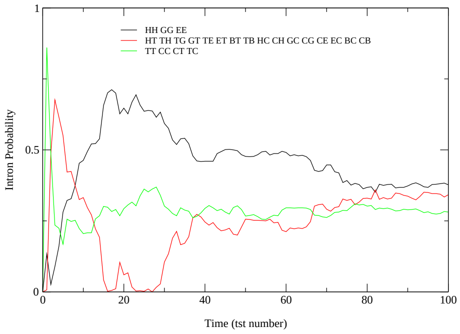


(b) without recombination

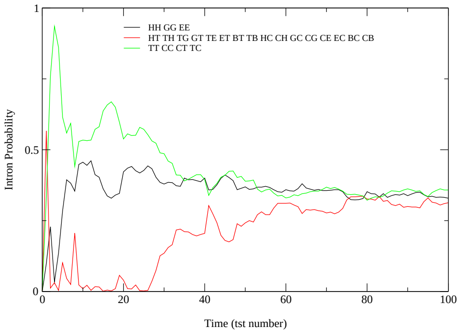
Figure 33: Comparison of the fitness development in a reactor simulation with exponential fitness function, with and without recombination. The reactor run shown in (a) was performed with the paramters displayed on table 1. In (b) the parameters are identical except for the recombination rate which was set to zero. 1LSE, startset mediumzscore2.



(a) Mediumzscore 2



(b) Mediumzscore 3



(c) Mediumzscore 4

Figure 34: Distribution of Introns in the lysozyme gene (1LSE) during a recombination reactor simulation,without recombination with exponential fitness ($\exp(z(S,\psi) - z(S_{wt},\psi))$, S_{wt} is the wild type sequence).

5 Conclusion and Outlook

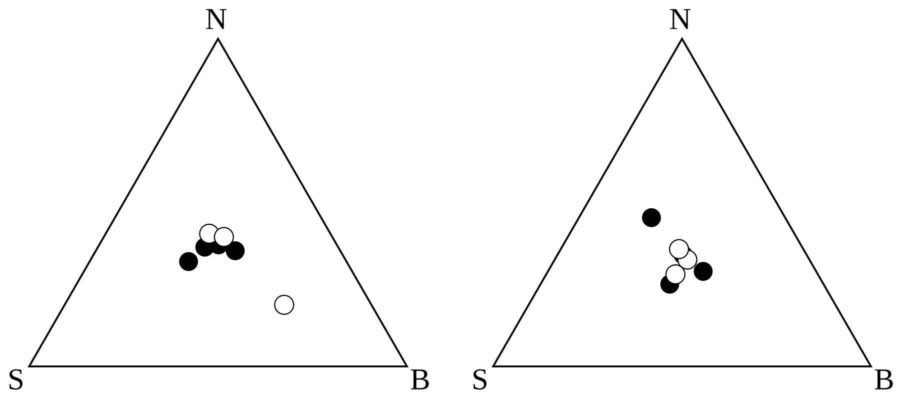
The Properties of Recombination In the introduction we assumed that recombination, without any respect to homology, will in the majority of cases act as a destructive force. We reasoned, that on average recombination events in the terminal regions of the biopolymer should have less effect than in the core parts and we therefore expected a funnel-like fitness versus recombination position curve. We were able to demonstrate this behaviour, both for recombination among a population of haploid species with an RNA genome (cf. 4.1.1), considering the RNA secondary structure as the phenotype and for a population of haploid species with a DNA genome coding for a single protein, were the zscore of the protein was the observed phenotype (cf. 4.1.2). We find apparent differences between the RNA secondary structure phenotype and the zscore phenotype in the steepness and smoothness of the curves measuring the impact of recombination on the expression of the phenotypes. Obviously a recombination event within a stack affects the formation of the RNA secondary structure severely even if the stack lies in the terminal parts of the RNA molecule. The discrepancy in the curves reflect the different fitness functions used in these experiments. RNA secondary structure formation is binary and singular, i.e. a nucleotide can either form a base pair or not, without any intermediate states and each nucleotide can participate only in at most one base pair. In contrast in the protein fitness function, each residue participates in a lot of interactions with different residues. The intensity of the interaction may be of any gradation. Due to this differences a recombination event in the RNA gene will either destroy the base pairing pattern or not, whereas a crossover in the protein gene, will most probably affect the phenotype, but not necessarily gravely.

We expected to find exceptions from the funnel-like trend at positions that separate structurally self-contained units, e.g. hinge regions in RNA, coil or turn regions in the protein structure or Gilbert's linker regions (cf. 2.3). In small RNA molecules, this is undoubtly true, see figure 7. After recombination at the hinge region more than one third of the recombinants retain the original mfe structure, in contrast to crossover points in the stacks (10% and 0%). In longer RNA molecules like the phenyl tRNA molecule, the effect of the recombination position cannot be read off in the fraction of molecules retaining the original mfe structure, which remains close to zero even at hinge positions, but in the minimal base pair pair distance of the recombinant's mfe structures to the original structure. One would expect to find similar positions in the protein experiment. Yet the zscore versus crossover position curve is much smoother than in the RNA case, with only small peaks. The more distinct peaks correlate with boundary regions in the protein secondary structure. Obviously only a small fraction of the recombinants, even at crossover positions one would expect to be linker regions, have better fitness than their parents.

We created a genetic algorithm based on recombination only, to investigate the effect of recombination on the variability in populations. The algorithm worked in the same way as the recombination statistics algorithm mentioned above, except that after each run of complete recombination, the best sequences of the offspring were elected to serve as parents for the next mating round. According to the formal analysis of recombination operators in chapter 2.2, recombination should homogenise population, reducing genetic variability. The analysis of Δz variance curves reveals indeed that the variance in the fitness decreases in the course of the run, which can be equated to a decreasing variability in the population. Another result of the analysis of the best Δz values of the algorithm, concerns the benefit of recombination. In this algorithm recombination plays not only a detrimental role, but produces with low frequency offspring with a higher fitness than its parents. However this effect depends, as in general any effect of recombination, on a sufficient variability in the population. We may conclude that recombination as an operator in an optimization process is only effective if it acts not on its own, but in combination with a second operator, which maintains a minimal level of heterogeneity.

Intron Placement and Recombination Introns could have been a means of genomes to adapt to the detrimental recombination. The dispute about the two classical theories of intron evolution, the introns early and introns late theory has now come to a paradox phase. Several firmed experiments confirm either the one or the other hypothesis, excluding the reverse one. As discussed in 2.3 our view of recombination could help to resolve this paradox conflict. On the one hand our view of recombination does not require a special time point for the evolution of introns, on the other hand it would explain the inhomogeneous distribution of introns within genes.

We studied the distribution of introns in genes under recombination in an evolutionary flow reactor simulation. Unfortunately we could not repeat the first encouraging result which suggested that introns form preferentially at borders of secondary structure elements. Further reactor runs yielded always different results, depending mainly on the initial population and on the seed of the random number generator. Analysis of the average fitness development in the reactor revealed that after a phase of continuous fitness gain, the average fitness oscillates over a large part of the fitness scale. We put this down to a too flat fitness function. Consequently we repeated the reactor runs with a steep fitness function ($\exp(z - z_{wt})$), in the hope to see more selection pressure for the modularisation of the protein genes. The new fitness function fulfilled its task to keep the average fitness fluctuations in the saturation phase small, but did not affect the intron distribution.



(a) Standard Fitness Function (b) Exponential Fitness Function

Figure 35: Comparison of the intron probabilities per secondary structure class at the 100th time step. S regions of defined secondary structure, B secondary structure borders and N regions of undefined secondary structure (cf. section 4.2). Filled circles symbolise runs with recombination (cf. table 1), open circles runs without recombination. An equal distribution of introns yields a disk in the center of the triangle.

In summary we cannot confirm that recombination causes a modularization of protein genes. Keeping up the original hypothesis that protein genes do modularize under recombination, one can explain the results by insufficient selection pressure for the modularization in our model. Eventually protein genes do have a tendency to modularize under recombination, but our experimental setup, makes it impossible to detect it. The reactor simulation is restricted in size, due to limitations in memory and computational effort. Modularization caused by recombination depends on a high variability among the population, because homologous recombination has no impact on the gene. Unfortunately we can keep only on the order of 1000 sequences in our reactor. We have shown in the genetic algorithm experiment (cf. 4.1.3) that recombination reduces the variability in the population. As a consequence the mutation rate in the reactor, which is responsible for variability, must be kept at a relatively high level, compared with the recombination frequency, thereby providing a constant level of variability. This technical constraint has an immediate consequence for the evolutionary process in the reactor: most offspring is created by mutation and only a small fraction by recombination. The chance to produce a recombinant with high fitness is low, especially because the crossover position is chosen at random. It might be possible that most of the recombinative offspring is washed out of reactor before it can produce offspring, due to its low fitness. This scenario results in a population, with most of its individuals having never been under the effect of recombination, consequently the distribution of introns would remain random, because point mutation does not interfere with introns.

A different explanation for the present results aims on the structure of the fitness function used. The fitness function is based on the assumption that structure defines function and that the ability of a sequence to fold into a specified structure can be estimated by rating each quadruple of residues with the likelihood of the quadruple. The concept of modularity in proteins lacks a proper physical bases. It is either the “optical” conclusion, that proteins are built from simple elements of similar geometry (secondary structure) or a notion of closeness (Gilbert’s idea of modules [23]). It is not clear whether this idea of modules has any thermodynamic relevance. The proteins used in the experiment were all of globular shape, the type of proteins for which the fitness function works best. Our concept of modularity would suggest, that it should be possible to replace an element of the protein, e.g. an entire alpha helix with a different alpha helix of the same length. However, in globular proteins, which are very compact, the interactions of the alpha helix’ side chains with their environment may be as important as the interactions inside the building block. Conseqeuently secondary structure elements might not act as modules at all.

Outlook The fitness function we used in our experiments is very limited in its applicability, which restricted our experiments to smaller protein folds without metal chelates or prosthetic groups. Protein structures which have evolved most probably through exon shuffling and should therefore have an inherent modularity in their structure could not be evaluated with the $\Delta z(x, \psi)$ function. A control experiment with such proteins would be of some importance to test, whether our computational setup can detect an already existing modularity.

It should be determined, whether the algebraic structure of the Δz fitness function is compatible with the notion of a fitness function that favours modularization. An algebraic description of modularize-able landscapes would be necessary to decide about this question. Koen Frenken and Luigi Marengo attempted recently to find a formalism for the decomposing a complex optimization problem into smaller subproblems [21]. They focused on the problem of nearly decomposability. A tradeoff between optimality of the solution and the performace to find it, could indeed play an important role in the evolution of modularity in biology. Alternatively one could investigate landscapes of the types $f(x) = f(x_1) + f(x_2)$ or $f(x) = f(x_1) \cdot f(x_2)$, where $f(x_1)$ is the contribution of one block and $f(x_2)$ of another block, via analysis of their amplitude spectra. However an exact Fourier decomposition is only possible for simple landscapes, landscapes which were used in the course of this work would be far to large.

A Abbreviations

Å	Ångstrøm (1 Å$= 10^{-10}$m)
ACR:	Ancient Conserved Region
cpDNA	Chloroplast DNA
CSTR:	Continuously Stirred Tank Reactor
DNA:	Desoxyribonucleic Acid
IE:	Introns Early Hypothesis
IL:	Introns Late Hypothesis
LCA:	Last Common Ancestor
Mbp	Mega Basepair
mRNA:	Messenger Ribinucleic Acid
mtDNA	Mitochondrial DNA
PDB:	Protein Data Bank
RNA:	Ribonucleic Acid
snRNP:	Small Nuclear Ribonucleoprotein
TPI:	Triosephosphate Isomerase
tRNA	Tranfer Ribonucleic Acid

B List of Figures

C List of Tables

D References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson. *Molecular Biology of the Cell*. Garland Press, 1994.

[2] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[3] S.M. Berget, C. Moore, and P.A. Sharp. Spliced segments at the 5' terminus of adenovirus-2 late mRNA. *Proc. Natl. Acad. Sci.*, 74:3171–3175, 1977.

[4] C. K. Biebricher, M. Eigen, and John S. McCaskill. Template-directed and template free RNA synthesis. *Journal of Molecular Biology*, 231:175–179, 1993.

[5] C. K. Biebricher and R. Luce. In vitro recombination and terminal elongation of RNA. *The EMBO Joernal*, 11:5129–5135, 1992.

[6] C. Brack and S. Tonegawa. Do genes-in-pieces imply protein-in-pieces? *Nature*, 273:267–268, 1977.

[7] R. Breathnach, J.L. Mandel, and P. Chambon. Ovalbumin gene is split in chicken. *Nature*, 270:314–319, 1977.

[8] R. Bürger. Evolution of genetic variability and the advantage of sex and recombination in changing environments. *Genetics*, 153:1055–1069, 1999.

[9] F. Cammas, J. Garnier, P. Chambon, and R. Losson. Correlation of the exon/intron organization to the conserved domains of the mouse transcriptional corepressor TIF 1β. *Gene*, 253:231–235, 2000.

[10] T. Cavalier-Smith. Intron phylogeny: a new hypothesis. *Trends in Genetics*, 7:145–148, 1991.

[11] P. Chambon. Split genes. *Scientific American*, 244:60–71, 1981.

[12] A. Chetverin. The puzzle of RNA recombination. *FEBS Letters*, 460:1–5, 1999.

[13] L.T. Chow, R.E. Gelinas, T.R. Broker, and R.J. Roberts. An amazing sequence arrangement at the 5' ends of adenovirus-2 messenger rna. *Cell*, 12:1–8, 1977.

[14] F. Crick. Split genes and rna splicing. *Science*, 204:264–271, 1979.

[15] J.E. Darnell. Implications of RNA. RNA splicing in evolution of eukaryotic cells. *Science*, 202:1257–1260, 1978.

[16] S.J. de Souza, M. Long, R.J. Klein, S. Roy, S. Lin, and W. Gilbert. Toward a resolution of the intron early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. *Gene*, 95:5094–5099, 1998.

[17] S.J. de Souza, M. Long, L. Schoenbach, S.W. Roy, and W. Gilbert. Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl. Acad. Sci. USA*, 93:14632–14636, 1996.

[18] S.J. de Souza, M. Long, L. Schoenbach, S.W. Roy, and W. Gilbert. The correlation between introns and the three-dimensional structure of proteins. *Gene*, 205:141–144, 1997.

[19] W.F. Doolittle. Genes in pieces: wre they ever together? *Nature*, 272:581–582, 1978.

[20] J.L. Ferat and F. Michel. Group II self-splicing introns in bacteria. *Nature*, 364:358–361, 1993.

[21] K. Frenken, L. Marengo, and M. Valente. Interdependencies, nearly-decomposability and adaptation. In T. Brenner, editor, *Computational Techniques for Modelling Learning in Economics.*, volume 11 of *Advances in Computational Economics*. Kluwer Academic Publishers, 1999.

[22] D.D.G. Gessler and X. Shizhong. Meiosis and the ecolution of recombination at low mutation rates. *Genetics*, 156:449–456, 2000.

[23] W. Gilbert. Why genes in pieces. *Nature*, 271:501, 1978.

[24] W. Gilbert. The rna world. *Nature*, 319:618, 1986.

[25] W. Gilbert. The exon theory of genes. *Cold Spring Habor Symp. Quant. Biol.*, 52:901–905, 1987.

[26] W. Gilbert and M. Glynias. On the ancient nature of introns. *Gene*, 135:137–144, 1993.

[27] P. Gitchoff and G. Wagner. Recombination induced hypergraphs: A new approch to mutation-recombination isomorphism. *Complexity*, 2:37–43, 1996.

[28] A.J.F. Griffiths, J.H. Miller, and D.T. Suzuki. *An Introduction to Genetic Analysis*. Unknown Books, 1996.

[29] I. Haruna and S. Spiegelman. Autocatalytic synthesis of a viral RNA in vitro. *Science*, 150:884–886, 1965.

[30] M. Hetzer, G. Wurzer, R.J. Schweyen, and M.W. Mueller. Trans-activation of group ii splicing by nuclear U5 snRNA. *Nature*, 386:417–420, 1997.

[31] J. Hey. Selfish genes and the origin of recombination. *Genetics*, 149:2089–2097, 1998.

[32] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Mh. Chem.*, 125:167–188, 1994.

[33] J.H. Holland. Building blocks, cohort genetic algorithms and hyperplane-defined functions. *Evolutionary Computation*, 8:373–391, 2000.

[34] J. Horn and D.E. Golberg. Genetic algorithm difficulty and the modality of fitness landscapes. In L. Darerell Whitley and Michael D. Vose, editors, *Foundations of Genetic Algorithms 3*, pages 243–269, San Francisco, CA, USA, 1995. Morgan Kaufmann.

[35] W. Jiang and A.V. Flannery. Correlation of the exon/intron organization to the secondary structures of the protease domain of mouse meprin α subunit. *Gene*, 189:65–71, 1997.

[36] V. Knoop, S. Kloska, and A. Brennicke. On the identification of group II introns in nucleotide sequence data. *Journal of Molecular Biology*, 242:389–396, 1994.

[37] A.S. Kondrashov. Classification of hypotheses on the advantage of aphimixis. *J. Hered.*, 84:372–384, 1993.

[38] J. Kwiatowski, M. Krawczyk, M. Kornacki, K. Bailey, and F.J. Ayala. Evidence against the exon theory of genes derived from the triosephosphate isomerase gene. *Proc. Natl.Acad. Sci*, 92:8503–8506, 1995.

[39] AM.M. Lambowitz and F. Belfort. Introns as mobile genetic elements. *Annu. Rev. Biochem*, 62:587–622, 1993.

[40] A. Landy. Dynamic, structural and regulatory aspects of lambda site-specific recombination. *Ann. Rev. Biochem.*, 58:913–949, 1989.

[41] T. Lenormand and S.P. Otto. The evolution of recombination in a heterogeneous environment. *Genetics*, 156:423–438, 2000.

[42] R.G. Lloyd and G.J. Sharples. Genetic analysis of recombination in prokaryotes. *Curr. Opin. Genet. Dev.*, 2:683–690, 1992.

[43] J.M. Jr Logsdon. The recent origins of spliceosomal introns revisited. *Current Opinion in Genetics & Developmet*, 8:637–648, 1998.

[44] J.M. Jr Logsdon, M.G. Tyshenko, C. Dixon, J.D. Jafari, V. Walker, and J.D. Palmer. Seven newly discovered intron positions in the triosephosphate isomerase gene: evidence for the introns-late theory. *Proc. Natl. Acad. Sci.*, 92:8507–8511, 1995.

[45] M. Long, C. Rosenberg, and W. Gilbert. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl. Acad. Sci*, 92:12495–12499, 1995.

[46] M. Marhcionni and W. Gilbert. The triosephosphate isomerase gene from maize: introns antendate the plant-animal divergence. *Cell*, 46:133–141, 1986.

[47] D.H. Mathew, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction

of rna secondary structure. *Journal of Molecular Biology*, 288:911–940, 1999.

[48] F. Michel and J.L. Ferat. Structure and activities of group II introns. *Annu. Rev. Biochem*, 64:435–461, 1995.

[49] Melanie Mitchell, John H. Holland, and Stephanie Forrest. When will a genetic algorithm outperform hill climbing. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 51–58. Morgan Kaufmann Publishers, Inc., 1994.

[50] K. Mizuuchi. Trnspositional recombination: mechanistic insights from studies of mu and other elements. *Ann. Rev. Biochemistry*, 61:1011–1051, 1992.

[51] A.V. Munishkin, L.A. Voronin, and A.B. Chetverin. An in vivo recombinant RNA capable of autokatalytic synthesis by q beta replicase. *Nature*, 333:473–475, 1988.

[52] R. Nussinov and Jacobson. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proc. Natl Acad. Sci.*, 77:6309–6313, 1980.

[53] R. Nussinov, G. Piecznik, and J.R. Griggs. Algorithms for loop matching. *SIAM J. Appl. Math*, 35:68–82, 1978.

[54] L. Patthy. Genome evolution and the evolution of exon shuffling - a review. *Gene*, 238:103–114, 1999.

[55] C. Rayssiguier, D.S. Thaler, and M. Radman. The barrier to recombination between escherichia coli and salmonella typhimurium is disrupted in mismatch repair mutants. *Nature*, 342:396–401, 1989.

[56] A.I. Roca and M.M. Cox. The recA protein: structure and function. *Crit. Rev. Biochem. Mol. Biol.*, 25:415–456, 1990.

[57] S.W. Roy, M. Noska, S.J. de Souza, and W. Gilbert. Centripedal modules and ancient introns. *Gene*, 238:85–91, 1999.

[58] A. Rzhetsky, F.J. Ayala, L.C. Hsu, C. Chang, and A. Yoshida. Exon/intron structure of aldehyde dehydrogenase genes supports the "introns-late" theory. *Proc. Natl. Acad. Sci*, 94:6820–6825, 1997.

[59] R. Saldanha, G. Mohr, F. Belfort, and AM. Lambowitz. Group I and group II introns. *FASEB J.*, 7:15–24, 1993.

[60] P.A. Sharp. Speculations on RNA splicing. *Cell*, 23:643–646, 1981.

[61] Manfred J. Sippl. Calculation of conformational ensembles from potentials of mean force — An approach to the knowledge-based prediction of local structures in globular proteins. *JMB*, 213:859–883, 1990.

[62] S. Spiegelman. An in vitro analysis of a replicating molecule. *American Scientist*, 55:63–68, 1967.

[63] Peter F. Stadler. Spectral landscape theory. In J. P. Crutchfield and P. Schuster, editors, *Evolutionary Dynamics—Exploring the Interplay of Selection, Neutrality, Accident, and Function*. Oxford University Press, 1999. in press.

[64] Peter F. Stadler and Günter P. Wagner. The algebraic theory of recombination spaces. *Evol. Comp.*, 5:241–275, 1998. Santa Fe Institute Preprint 96-07-046.

[65] P.F. Stadler, R. Seitz, and G.P. Wagner. Population dependent fourier decomposition of fitness landscapes over recombination spaces: Evolveability of complex characters. *Bulletin of Mathematical Biology*, 00:1–30, 1999.

[66] A. Stoltzfus, D.F. Spencer, M. Zuker, J.M. Logsdon, and W.F. Doolittle. Testig the exon theory of genes: The evidence from protein structure. *Science*, 265:202–207, 1994.

[67] D. Straus and W. Gilbert. Genetic engeneering in the precambrian: structure of the chicken triosephosphate isomerase gene. *Mol. Cell Biol.*, 5:3497–3506, 1985.

[68] Günter P. Wagner and Peter F. Stadler. Complex adaptations and the structure of recombination spaces. In Chrystopher Nehaniv and Misami Ito, editors, *Algebraic Engineering*, pages 96–115, Singapore, 1999.

World Scientific. (Proceedings of the Conference on Semi-Groups and Algebraic Engineering, University of Aizu, Japan); Santa Fe Institute Preprint 97-03-029.

[69] G Weberndorfer. Empirical protein potentials from delauney tesselation. Master's thesis, University of Vienna, 1999.

[70] G. Weberndorfer, I.L. Hofacker, and P.F. Stadler. An efficient potential for protein sequence design. In *Giegerich et al. (eds) Computer Science in Biology [GCB'99 Proceedings]*, pages 73–79, 1999.

[71] D. Wolpert and W. MacReady. No free lunch theorems for optimization, 1996.

[72] A.J. Zaug and T.R. Cech. The intervening sequence of RNA of tetrahymena is an enzyme. *Science*, 231:470–475, 1986.

[73] L.A. Zhivotovsky, M.W. Feldman, and F.B. Christiansen. Evolution of recombination among multiple selected loci: A generalized reduction principle. *Proc. Natl. Acad. Sci.*, 91:1079–1083, 1994.

[74] M. Zuker and P. Stiegler. Optimal computerfolding of large RNA sequences using thermodynamic and auxiliary information. *Nucl. Acid Res*, 9:133–148, 1981.

Curriculum vitae

Jörg Hackermüller
* 3. Jänner 1976 in Steyr, Oberösterreich

Ausbildung

1981 – 1985	Volksschule Ulmerfeld Hausmening
1986 – 1994	Bundesreagymnasium Waidhofen/Ybbs
Juni, 1994	Matura, mit Auszeichnung
1994 – 2001	Chemiestudium an der Universität Wien
Winter 1999	Auslandsaufenthalt am Karolinska Institut und Smittskyddsinstitutet, beide Stockholm, Schweden
10/2000 – 2/2001	Diplomarbeit am Institut fuer Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien, bei Prof. Peter Stadler in der Gruppe von Prof. Peter Schuster.
seit 2/2000	Zivildienst, Arbeitersamariterbund Österreichs